Benign Overfitting in Multiclass Classification: All Roads Lead to Interpolation

Ke Wang* Vidya Muthukumar† Christos Thrampoulidis°

*Department of Statistics and Applied Probability, University of California Santa Barbara
†Electrical and Computer Engineering & Industrial and Systems Engineering, Georgia Institute of Technology
°Department of Electrical and Computer Engineering, University of British Columbia

Abstract

The literature on "benign overfitting" in overparameterized models has been mostly restricted to regression or binary classification; however, modern machine learning operates in the multiclass setting. Motivated by this discrepancy, we study benign overfitting in multiclass linear classification. Specifically, we consider the following training algorithms on separable data: (i) empirical risk minimization (ERM) with cross-entropy loss, which converges to the multiclass support vector machine (SVM) solution; (ii) ERM with least-squares loss, which converges to the min-norm interpolating (MNI) solution; and, (iii) the one-vs-all SVM classifier. First, we provide a simple sufficient deterministic condition under which *all* three algorithms lead to classifiers that interpolate the training data and have equal accuracy. When the data is generated from Gaussian mixtures or a multinomial logistic model, this condition holds under high enough effective overparameterization. We also show that this sufficient condition is satisfied under "neural collapse", a phenomenon that is observed in training deep neural networks. Second, we derive novel bounds on the accuracy of the MNI classifier, thereby showing that all three training algorithms lead to benign overfitting under sufficient overparameterization. Ultimately, our analysis shows that good generalization is possible for SVM solutions beyond the realm in which typical margin-based bounds apply.

1 Introduction

Modern deep neural networks are overparameterized (high-dimensional) with respect to the amount of training data. Consequently, they achieve zero training error even on noisy training data, yet generalize well on test data [ZBH+17]. Recent mathematical analysis has shown that fitting of noise in regression tasks can in fact be relatively benign for linear models that are sufficiently high-dimensional [BLLT20, BHX20, HMRT19, MVSS20, KLS20]. These analyses do not directly extend to classification, which requires separate treatment. In fact, recent progress on sharp analysis of interpolating binary classifiers [MNS+21, CL21, WT21, CGB21] revealed high-dimensional regimes in which binary classification generalizes well, but the corresponding regression task does *not* work and/or the success *cannot* be predicted by classical margin-based bounds [SFBL98, BM03].

In an important separate development, these same high-dimensional regimes admit an equivalence of loss functions used for optimization at training time. The support vector machine (SVM), which arises from minimizing the logistic loss using gradient descent [SHN+18, JT19], was recently shown to satisfy a high-probability equivalence to interpolation, which arises from minimizing the squared loss [MNS+21, HMX21]. This equivalence suggests that interpolation is ubiquitous in very overparameterized settings, and can arise naturally as a consequence of the optimization procedure even when this is not explicitly encoded or intended. Moreover, this equivalence to interpolation and corresponding analysis implies that the SVM can generalize even in regimes where classical learning theory bounds are not predictive. In the logistic model case [MNS+21] and Gaussian binary mixture model case [CL21, WT21, CGB21],

*Primary correspondence to: kewang01@ucsb.edu. CT is also affiliated with the Department of Electrical and Computer Engineering, University of California, Santa Barbara.

it is shown that good generalization of the SVM is possible beyond the realm in which classical margin-based bounds apply. These analyses lend theoretical grounding to the surprising hypothesis that *squared loss can be equivalent to, or possibly even superior*, to the cross-entropy loss for classification tasks. Ryan Rifkin provided empirical support for this hypothesis on kernel machines [Rif02, RK04]; more recently, corresponding empirical evidence has been provided for state-of-the-art deep neural networks [HB20, PL20a].

These perspectives have thus far been limited to regression and *binary* classification settings. In contrast, most success stories and surprising new phenomena of modern machine learning have been recorded in *multiclass* classification settings, which appear naturally in a host of applications that demand the ability to automatically distinguish between large numbers of different classes. For example, the popular ImageNet dataset [RDS+15] contains on the order of 1000 classes. Whether a) good generalization beyond effectively low-dimensional regimes where margin-based bounds are predictive is possible, and b) equivalence of squared loss and cross-entropy loss holds in multiclass settings remained open problems.

This paper makes significant progress towards a complete understanding of the optimization and generalization properties of high-dimensional linear multiclass classification, both for unconditional Gaussian covariates (where labels are generated via a multinomial logistic model), and Gaussian mixture models. Our contributions are listed in more detail below.

1.1 Our Contributions

• We establish a *deterministic* sufficient condition under which the multiclass SVM solution has a very simple and symmetric structure: it is identical to the solution of a One-vs-All (OvA) SVM classifier that uses a *simplex-type* encoding for the labels (unlike the classical one-hot encoding). Moreover, the constraints at both solutions are active. Geometrically, this means that all data points are support vectors, and *they interpolate the simplex-encoding vector representation of the labels.* See Figure 2 for a numerical illustration confirming our finding.

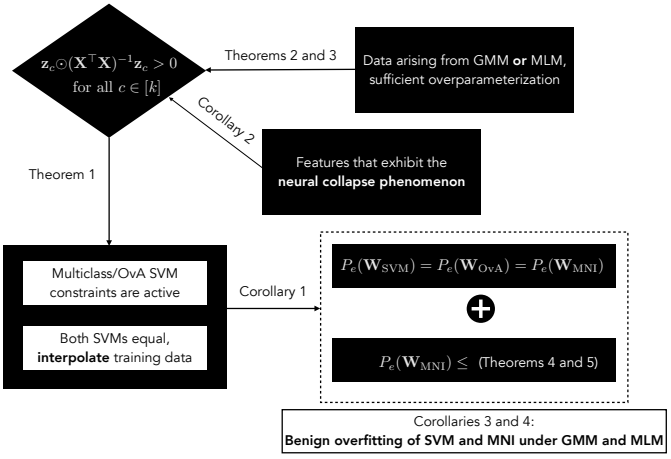


Figure 1: Contributions and organization.

• This implies a surprising equivalence between traditionally different formulations of multiclass SVM, which in turn are equivalent to the minimum-norm interpolating (MNI) classifier on the one-hot label vectors. Thus, we show that the outcomes of training with cross-entropy (CE) loss and squared loss are identical in terms of classification error.

• Next, for data following a Gaussian-mixtures model (GMM) or a Multinomial logistic model (MLM), we show that the above sufficient condition is satisfied with high-probability under sufficient "effective" overparameterization. Our sufficient conditions are non-asymptotic and are characterized in terms of the data dimension, the number of classes, and functionals of the data covariance matrix. Our numerical results show excellent agreement with our theoretical findings. We also show that the sufficient condition of equivalence of CE and squared losses is satisfied when the "neural collapse" phenomenon occurs [PHD20].

• Finally, we provide novel non-asymptotic bounds on the error of the MNI classifier for data generated either from the GMM or the MLM, and identify sufficient conditions under which benign overfitting occurs. A direct outcome of our results is that benign overfitting occurs under these conditions regardless of whether the cross-entropy loss or squared loss is used during training.

Figure 1 describes our contributions and their implications through a flowchart. To the best of our

knowledge, these are the first results characterizing a) equivalence of loss functions, and b) generalization of interpolating solutions in the multiclass setting. The multiclass setting poses several challenges over and above the recently studied binary case. When presenting our results in later sections, we discuss in detail how our analysis circumvents these challenges.

1.2 Related Work

Multiclass classification and the impact of training loss functions There is a classical body of work on algorithms for multiclass classification, e.g. [WW98, BB99, DB95, CS02, LLW04] and several empirical studies of their comparative performance [RK04, FÖ2, ASS01] (also see [HYS16, GCOZ17, KS18, BEH20, DCO20, HB20, PL20a] for recent such studies in the context of deep nets). Many of these (e.g. [RK04, HB20, BEH20]) have found that least-squares minimization yields competitive test classification performance to cross-entropy minimization. *Our proof of equivalence of the SVM and MNI solutions under sufficient overparameterization provides theoretical support for this line of work.* This is a consequence of the implicit bias of gradient descent run on the CE and squared losses leading to the multiclass SVM [SHN+18, JT19] and MNI [EHN96] respectively. Numerous classical works investigated consistency [Zha04, LLW04, TB07, PGS13, PS16] and finite-sample behavior, e.g., [KP02, CKMY16, LDBK15, Mau16, LDZK19, MR16] of multiclass classification algorithms in the underparameterized regime. In contrast, our primary focus lies in the highly overparameterized regime, where conventional techniques of uniform convergence are inadequate. In Section IV, we elaborate on why classical training-data-dependent bounds based on margin or Rademacher complexity, are insufficient in this regime and cannot yield conclusions about benign overfitting. Recently, in [AGL21], the authors have studied the problem of feature selection in high-dimensional multiclass classification, identifying an intriguing phase transition as the number of classes increases with dimensions. Our work differs from theirs in that our bounds are relevant to CE minimization without explicit regularization, whereas [AGL21] focuses on CE loss minimization with sparsity penalties to achieve feature selection.

Binary classification error analyses in overparameterized regime The recent wave of analyses of the minimum-ℓ_2-norm interpolator (MNI) in high-dimensional linear regression (beginning with [BLLT20, BHX20, HMRT19, MVSS20, KLS20]) prompted researchers to consider to what extent the phenomena of benign overfitting and double descent [BHMM19, GJS+20] can be proven to occur in classification tasks. Even the binary classification setting turns out to be significantly more challenging to study owing to the discontinuity of the 0-1 test loss function. Sharp asymptotic formulas for the generalization error of binary classification algorithms in the linear high-dimensional regime have been derived in several recent works [Hua17, SC19, MLC19, SAH19, TPT20, TPT21, DKT21, MRSY19, KA21, LS20, SAH20, AKLZ20, Lol20, DL20]. These formulas are solutions to complicated nonlinear systems of equations that typically do not admit closed-form expressions. A separate line of work provides non-asymptotic error bounds for both the MNI classifier and the SVM classifier [CL21, MNS+21, WT21, CGB21]; in particular, [MNS+21] analyzed the SVM in a Gaussian covariates model by explicitly connecting its solution to the MNI solution. Subsequently, [WT21] also took this route to analyze the SVM and MNI in mixture models, and even more recently, [CGB21] provided extensions of this result to sub-Gaussian mixtures. While these non-asymptotic analyses are only sharp in their dependences on the sample size n and the data dimension p, they provide closed-form generalization expressions in terms of easily interpretable summary statistics. Interestingly, these results imply good generalization of the SVM beyond the regime in which margin-based bounds are predictive. Specifically, [MNS+21] identifies a separating regime for Gaussian covariates in which corresponding regression tasks would not generalize. In the Gaussian mixture model, margin-based bounds [SFBL98, BM03] (as well as corresponding recently derived mistake bounds on interpolating classifiers [LR21]) would require the intrinsic signal-to-noise-ratio (SNR) to scale at least as $\omega(p^{1/2})$ for good generalization; however, the analyses of [WT21, CGB21] show that good generalization is possible for significantly lower SNR scaling as $\omega(p^{1/4})$. The above error analyses are specialized to the binary case, where closed-form error expressions are easy to derive [MNS+21]. The only related work applicable to the multiclass case is [TOS20], which also highlights the numerous challenges of obtaining a sharp error analysis in multiclass settings. Specifically, [TOS20] derived sharp generalization formulas for multiclass least-squares in

underparameterized settings; extensions to the overparameterized regime and other losses beyond least-squares remained open. Finally, [KT21] recently derived sharp phase-transition thresholds for the feasibility of OvA-SVM on multiclass Gaussian mixture data in the linear high-dimensional regime. However, this does not address the more challenging multiclass-SVM that we investigate here. To summarize, our paper presents the first generalization bounds for the multiclass-SVM classifier that establish conditions for benign overfitting in the high-dimensional regime. In the process, we establish a connection between multiclass-SVM and multi-class MNI, which poses unique challenges due to the non-uniqueness of defining support vectors in multiclass settings. Our work highlights the richness of the multiclass setting compared to the binary setting, as we demonstrate the equivalence not only for GMM and MLM data but also for data following a simplex equiangular tight-frame (ETF) structure. This geometry structure is only relevant in multiclass settings and arises when training deep-net classifiers with CE loss beyond the zero-training error regime [PHD20].

Other SVM analyses The number of support vectors in the *binary* SVM has been characterized in low-dimensional separable and non-separable settings [DOS99, BG01, MO05] and scenarios have been identified in which there is a vanishing fraction of support vectors, as this implies good generalization[1] via PAC-Bayes sample compression bounds [Vap13, GHST05, GLL+11]. In the highly overparameterized regime that we consider, perhaps surprisingly, the opposite behavior occurs: *all training points become support vectors with high probability* [DOS99, BG01, MO05, MNS+21, HMX21]. In particular, [HMX21] provided sharp non-asymptotic sufficient conditions for this phenomenon for both isotropic and anisotropic settings. The techniques in [MNS+21, HMX21] are highly specialized to the binary SVM and its dual, where a simple complementary slackness condition directly implies the property of interpolation. In contrast, the complementary slackness condition for the case of multiclass SVM *does not* directly imply interpolation; in fact, the operational meaning of "all training points becoming support vectors" is unclear in the multiclass SVM. *Our proof of deterministic equivalence goes beyond the complementary slackness condition and uncovers a surprising symmetric structure by showing equivalence of multiclass SVM to a simplex-type OvA classifier.* The simplex equiangular tight frame structure that we uncover is somewhat reminiscent of the recently observed neural collapse phenomenon in deep neural networks [PHD20]; indeed, *Section 3.3 shows an explicit connection between our deterministic equivalence condition and the neural collapse phenomenon.* Further, [MNS+21, HMX21] focus on proving deterministic conditions for equivalence in the case of labels generated from covariates; the mixture model case (where covariates are generated from labels) turns out to be significantly more involved due to the *anisotropic* data covariance matrix resulting from even from isotropic noise covariance [WT21, CGB21]. As we explain further in Section 3.2, the mean vectors of the mixture model introduce an additional rank-k component that complicates the analysis and requires new ideas.

1.3 Organization

The paper is organized as follows. Section 2 describes the problem setting and sets up notation. Section 3 presents our main results on the equivalence between the multiclass SVM and MNI solutions for two data models: the Gaussian mixture model (GMM) and the multinomial logistic model (MLM). In the same section, we also show the equivalence under the Neural Collapse phenomenon. Section 4 presents our error analysis of the MNI solution (and, by our proved equivalence, the multiclass SVM) for the GMM and the MLM, and Section 5 presents consequent conditions for benign overfitting of multiclass classification. Finally, Section 6 presents proofs of our main results; auxiliary proofs are deferred to the appendices. Please refer to the table of contents (before the appendices) for a more detailed decomposition of results and proofs.

Notation For a vector $\mathbf{v} \in \mathbb{R}^p$, let $\|\mathbf{v}\|_2 = \sqrt{\sum_{i=1}^p v_i^2}$, $\|\mathbf{v}\|_1 = \sum_{i=1}^p |v_i|$, $\|\mathbf{v}\|_\infty = \max_i\{|v_i|\}$. $\mathbf{v} > \mathbf{0}$ is interpreted elementwise. $\mathbf{1}_m$ / $\mathbf{0}_m$ denote the all-ones / all-zeros vectors of dimension m

[1] In this context, the fact that [MNS+21, WT21, CGB21] provide good generalization bounds in the regime where support vectors proliferate is particularly surprising. In conventional wisdom, a proliferation of support vectors was associated with overfitting but this turns out to not be the case here.

and $\mathbf{e}_i$ denotes the i-th standard basis vector. For a matrix $\mathbf{M}$, $\|\mathbf{M}\|_2$ denotes its $2 \to 2$ operator norm and $\|\mathbf{M}\|_F$ denotes the Frobenius norm. $\odot$ denotes the Hadamard product. $[n]$ denotes the set $\{1, 2, ..., n\}$. We also use standard "Big O" notations $\Theta(\cdot)$, $\omega(\cdot)$, e.g. see [CLRS09, Chapter 3]. Finally, we write $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the (multivariate) Gaussian distribution of mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and, $Q(x) = \mathbb{P}(Z > x),\ Z \sim \mathcal{N}(0, 1)$ for the Q-function of a standard normal. Throughout, constants refer to strictly positive numbers that do not depend on the problem dimensions n or p.

2 Problem setting

We consider the multiclass classification problem with k classes. Let $\mathbf{x} \in \mathbb{R}^p$ denote the feature vector and $y \in [k]$ represent the class label associated with one of the k classes. We assume that the training data has n feature/label pairs $\{\mathbf{x}_i, y_i\}_{i=1}^n$. We focus on the overparameterized regime, i.e. $p > Cn$, and we will frequently consider $p \gg n$. For convenience, we express the labels using the one-hot coding vector $\mathbf{y}_i \in \mathbb{R}^k$, where only the y_i-th entry of $\mathbf{y}_i$ is 1 and all other entries are zero, i.e. $\mathbf{y}_i = \boldsymbol{e}_{y_i}$. With this notation, the feature and label matrices are given in compact form as follows: $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n\end{bmatrix} \in \mathbb{R}^{p \times n}$ and $\mathbf{Y} = \begin{bmatrix}\mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_n\end{bmatrix} = \begin{bmatrix}\mathbf{v}_1 & \mathbf{v}_2 & \cdots \mathbf{v}_k\end{bmatrix}^T \in \mathbb{R}^{k \times n}$, where we have defined $\mathbf{v}_c \in \mathbb{R}^n, c \in [k]$ to denote the c-th row of the matrix $\mathbf{Y}$.

2.1 Data models

We assume that the data pairs $\{\mathbf{x}_i, y_i\}_{i=1}^n$ are independently and identically distributed (IID). We will consider two models for the distribution of $(\mathbf{x}, y)$. For both models, we define the mean vectors $\{\boldsymbol{\mu}_j\}_{j=1}^k \in \mathbb{R}^p$, and the mean matrix is given by $\mathbf{M} := \begin{bmatrix}\boldsymbol{\mu}_1 & \boldsymbol{\mu}_2 & \cdots & \boldsymbol{\mu}_k\end{bmatrix} \in \mathbb{R}^{p \times k}$.

Gaussian Mixture Model (GMM) In this model, the mean vector $\boldsymbol{\mu}_i$ represents the conditional mean vector for the i-th class. Specifically, each observation $(\mathbf{x}_i, y_i)$ belongs to to class $c \in [k]$ with probability π_c and conditional on the label y_i, $\mathbf{x}_i$ follows a multivariate Gaussian distribution. In summary, we have

$$\mathbb{P}(y = c) = \pi_c \ \text{ and } \ \mathbf{x} = \boldsymbol{\mu}_y + \mathbf{q},\ \mathbf{q} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{1}$$

In this work, we focus on the isotropic case $\boldsymbol{\Sigma} = \mathbf{I}_p$. Our analysis can likely be extended to the more general anisotropic case, but we leave this to future work.

Multinomial Logit Model (MLM) In this model, the feature vector $\mathbf{x} \in \mathbb{R}^p$ is distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, and the conditional density of the class label y is given by the soft-max function. Specifically, we have

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \ \text{ and } \ \mathbb{P}(y = c|\mathbf{x}) = \frac{\exp(\boldsymbol{\mu}_c^T \mathbf{x})}{\sum_{j \in [k]} \exp(\boldsymbol{\mu}_j^T \mathbf{x})}. \tag{2}$$

For this model, we analyze both the isotropic and anisotropic cases.

2.2 Data separability

We consider linear classifiers parameterized by $\mathbf{W} = \begin{bmatrix}\mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_k\end{bmatrix}^T \in \mathbb{R}^{k \times p}$. Given input feature vector $\mathbf{x}$, the classifier is a function that maps $\mathbf{x}$ into an output of k via $\mathbf{x} \mapsto \mathbf{W}\mathbf{x} \in \mathbb{R}^k$ (for simplicity, we ignore the bias term throughout). We will operate in a regime where the training data are linearly separable. In multiclass settings, there exist multiple notions of separability. Here, we focus on (i) multiclass separability (also called k-class separability) (ii) one-vs-all (OvA) separability, and, recall their definitions below.

Definition 1 (multiclass and OvA separability)**.** *The dataset $\{\mathbf{x}_i, y_i\}_{i \in [n]}$ is multiclass linearly separable when*

$$\exists \mathbf{W} \ : \ (\mathbf{w}_{y_i} - \mathbf{w}_c)^T \mathbf{x}_i \geq 1,\ \forall c \neq y_i, c \in [k],\ \textit{and}\ \forall i \in [n]. \tag{3}$$

The dataset is one-vs-all (OvA) separable when

$$\exists \mathbf{W} \,:\, \mathbf{w}_c^T \mathbf{x}_i \begin{cases} \geq 1 \ \textit{if } y_i = c \\ \leq -1 \ \textit{if } y_i \neq c \end{cases}, \forall c \in [k], \ \textit{and } \forall i \in [n]. \tag{4}$$

Under both data models of the previous section (i.e. GMM and MLM), we have $\text{rank}(\mathbf{X}) = \text{n}$ almost surely in the overparameterized regime $p > n$. This directly implies OvA separability. It turns out that OvA separability implies multiclass separability, but not vice versa (see [BM94] for a counterexample).

2.3 Classification error

Consider a linear classifier $\widehat{\mathbf{W}}$ and a fresh sample $(\mathbf{x}, y)$ generated following the same distribution as the training data. As is standard, we predict $\hat{y}$ by a "winner takes it all strategy", i.e. $\hat{y} = \arg\max_{j \in [k]} \widehat{\mathbf{w}}_j^T \mathbf{x}$. Then, the classification error conditioned on the true label being c, which we refer to as the *class-wise classification error*, is defined as

$$\mathbb{P}_{e|c} := \mathbb{P}(\hat{y} \neq y | y = c) = \mathbb{P}(\widehat{\mathbf{w}}_c^T \mathbf{x} \leq \max_{j \neq c} \widehat{\mathbf{w}}_j^T \mathbf{x}). \tag{5}$$

In turn, the *total classification error* is defined as

$$\mathbb{P}_e := \mathbb{P}(\hat{y} \neq y) = \mathbb{P}(\arg\max_{j \in [k]} \widehat{\mathbf{w}}_j^T \mathbf{x} \neq y) = \mathbb{P}(\widehat{\mathbf{w}}_y^T \mathbf{x} \leq \max_{j \neq y} \widehat{\mathbf{w}}_j^T \mathbf{x}). \tag{6}$$

2.4 Classification algorithms

Next, we review several different training strategies for which we characterize the total/class-wise classification error in this paper.

Multiclass SVM Consider training $\mathbf{W}$ by minimizing the cross-entropy (CE) loss

$$\mathcal{L}(\mathbf{W}) := -\log\left(\frac{e^{\mathbf{w}_{y_i}^T \mathbf{x}_i}}{\sum_{c \in [k]} e^{\mathbf{w}_c^T \mathbf{x}_i}}\right)$$

with the gradient descent algorithm (with constant step size η). In the separable regime, the CE loss $\mathcal{L}(\mathbf{W})$ can be driven to zero. Moreover, [SHN+18, Thm. 7] showed that the normalized iterates $\{\mathbf{W}^t\}_{t \geq 1}$ converge as

$$\lim_{t \to \infty} \left\| \frac{\mathbf{W}^t}{\log t} - \mathbf{W}_{\text{SVM}} \right\|_F = 0,$$

where $\mathbf{W}_{\text{SVM}}$ is the solution of the *multiclass SVM* [WW98] given by

$$\mathbf{W}_{\text{SVM}} := \arg\min_{\mathbf{W}} \|\mathbf{W}\|_F \quad \text{sub. to} \quad (\mathbf{w}_{y_i} - \mathbf{w}_c)^T \mathbf{x}_i \geq 1, \ \forall i \in [n], c \in [k] \text{ s.t. } c \neq y_i. \tag{7}$$

It is important to note that the normalizing factor $\log t$ here does *not* depend on the class label; hence, in the limit of GD iterations, the solution $\mathbf{W}^t$ decides the same label as multiclass SVM for any test sample.

One-vs-all SVM In contrast to Equation (7), which optimizes the hyperplanes $\{\mathbf{w}_c\}_{c \in [k]}$ jointly, the one-vs-all (OvA)-SVM classifier solves k separable optimization problems that maximize the margin of each class with respect to all the rest. Concretely, the OvA-SVM solves the following optimization problem for all $c \in [k]$:

$$\mathbf{w}_{\text{OvA},c} := \arg\min_{\mathbf{w}} \|\mathbf{w}\|_2 \quad \text{sub. to} \quad \mathbf{w}^T \mathbf{x}_i \begin{cases} \geq 1, & \text{if } \mathbf{y}_i = c, \\ \leq -1, & \text{if } \mathbf{y}_i \neq c, \end{cases} \quad \forall i \in [n]. \tag{8}$$

In general, the solutions to Equations (7) and (8) are different. While the OvA-SVM does not have an obvious connection to any training loss function, its relevance will become clear in Section 3. Perhaps surprisingly, we will prove that in the highly overparameterized regime the multiclass SVM solution is identical to a slight variant of (8).

Min-norm interpolating (MNI) classifier An alternative to the CE loss is the square loss $\mathcal{L}(\mathbf{W}) := \frac{1}{2n}\|\mathbf{Y} - \mathbf{W}\mathbf{X}\|_2^2 = \frac{1}{2n}\sum_{i=1}^n \|\mathbf{W}\mathbf{x}_i - \mathbf{y}_i\|_2^2$. Since the square loss is tailored to regression, it might appear that the CE loss is more appropriate for classification. Perhaps surprisingly, one of the main messages of this paper is that under sufficient effective overparameterization the two losses actually have equivalent performance. Our results lend theoretical support to empirical observations of competitive classification accuracy between the square loss and CE loss in practice [Rif02, HB20, PL20a].

Towards showing this, we note that when the linear model is overparameterized (i.e. $p > n$) and assuming rank($\mathbf{X}$) = n (e.g this holds almost surely under both the GMM and MLM), the data can be linearly interpolated, i.e. the square-loss can be driven to zero. Then, it is well-known [EHN96] that gradient descent with sufficiently small step size and appropriate initialization converges to the minimum-norm -interpolating (MNI) solution, given by:

$$\mathbf{W}_{\text{MNI}} := \arg\min_{\mathbf{W}} \|\mathbf{W}\|_F, \quad \text{sub. to} \quad \mathbf{X}^T\mathbf{w}_c = \mathbf{v}_c, \forall c \in [k]. \tag{9}$$

Since $\mathbf{X}^T\mathbf{X}$ is invertible, the MNI solution is given in closed form as $\mathbf{W}_{\text{MNI}}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y}^T$. From here on, we refer to (9) as the MNI classifier.

3 Equivalence of solutions and geometry of support vectors

In this section, we show the equivalence of the solutions of the three classifiers defined above in certain high-dimensional regimes.

3.1 A key deterministic condition

We first establish a key deterministic property of SVM that holds for *generic* multiclass datasets $(\mathbf{X}, \mathbf{Y})$ (i.e. not necessarily generated by either the GMM or MLM). Specifically, Theorem 1 below derives a sufficient condition (cf. (12)) under which the multiclass SVM solution has a surprisingly simple structure. First, the constraints are *all* active at the optima (cf. (13)). Second, and perhaps more interestingly, this happens in a very specific way; the feature vectors interpolate a simplex representation of the multiclass labels, as specified below:

$$\hat{\mathbf{w}}_c^T\mathbf{x}_i = z_{ci} := \begin{cases} \frac{k-1}{k} & , \ c = y_i \\ -\frac{1}{k} & , \ c \neq y_i \end{cases} \quad \text{for all } i \in [n], c \in [k]. \tag{10}$$

To interpret this, define an adjusted k-dimensional label vector $\widetilde{\mathbf{y}}_i := [z_{1i}, z_{2i}, \ldots, z_{ki}]^T$ for each training sample $i \in [n]$. This can be understood as a k-dimensional vector encoding of the original label y_i that is different from the classical one-hot encoding representation $\mathbf{y}_i$; in particular, it has entries either $-1/k$ or $1 - 1/k$ (rather than 0 or 1). We call this new representation a *simplex representation*, based on the following observation. Consider k data points that each belong to a different class $1, \ldots, k$, and their corresponding vector representations $\widetilde{\mathbf{y}}_1, \ldots, \widetilde{\mathbf{y}}_k$. Then, it is easy to verify that the vectors $\{\mathbf{0}, \widetilde{\mathbf{y}}_1, \ldots, \widetilde{\mathbf{y}}_k\}$ are affinely independent; hence, they form the vectices of a k-simplex.

Theorem 1. *For a multiclass separable dataset with feature matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$ and label matrix $\mathbf{Y} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k]^T \in \mathbb{R}^{k\times n}$, denote by $\mathbf{W}_{SVM} = [\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_k]^T$ the multiclass SVM solution of (7). For each class $c \in [k]$ define vectors $\mathbf{z}_c \in \mathbb{R}^n$ such that*

$$\mathbf{z}_c = \mathbf{v}_c - \frac{1}{k}\mathbf{1}_n, \ c \in [k]. \tag{11}$$

Let $(\mathbf{X}^T\mathbf{X})^+$ be the Moore-Penrose generalized inverse[2] of the Gram matrix $\mathbf{X}^T\mathbf{X}$ and assume that the following condition holds

$$\mathbf{z}_c \odot (\mathbf{X}^T\mathbf{X})^+\mathbf{z}_c > \mathbf{0}, \quad \forall c \in [k]. \tag{12}$$

[2]Most of the regimes that we study are ultra-high-dimensional (i.e. $p \gg n$), and so $\mathbf{X}^T\mathbf{X}$ is invertible with high probability. Consequently, $(\mathbf{X}^T\mathbf{X})^+$ can be replaced by $(\mathbf{X}^T\mathbf{X})^{-1}$ in these cases.

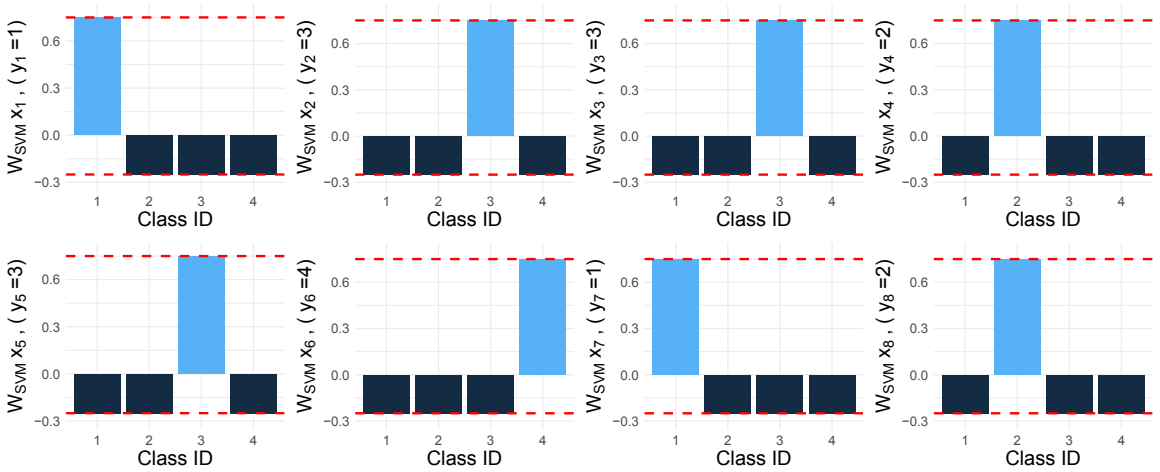


Figure 2: Inner products $\mathbf{W}_{\mathrm{SVM}}\mathbf{x}_c \in \mathbb{R}^4$ for features $\mathbf{x}_i$ that each belongs to the c-th class for $c \in [k]$ and $k = 4$ total classes. The red lines correspond to the values $(k-1)/k = 3/4$ and $-1/k = -1/4$ of the simplex encoding described in Theorem 1. Observe that the inner products $\mathbf{W}_{\mathrm{SVM}}\mathbf{x}_c$ match with these values, that is, Equation (10) holds.

Then, the SVM solution $\mathbf{W}_{SVM}$ *is such that all the constraints in* (7) *are active. That is,*

$$(\hat{\mathbf{w}}_{y_i} - \hat{\mathbf{w}}_c)^T \mathbf{x}_i = 1, \ \forall c \neq y_i, c \in [k], \ \textit{and} \ \ \forall i \in [n]. \tag{13}$$

Moreover, the features interpolate the simplex representation. That is,

$$\mathbf{X}^T \hat{\mathbf{w}}_c = \mathbf{z}_c, \ \forall c \in [k]. \tag{14}$$

For $k = 2$ classes, it can be easily verified that Equation (12) reduces to the condition in Equation (22) of [MNS+21] for the binary SVM. Compared to the binary setting, the conclusion for the multiclass case is richer: provided that Equation (12) holds, we show that not only are all data points support vectors, but also, they satisfy a set of simplex OvA-type constraints as elaborated above. The proof of Equation (14) is particularly subtle and involved: unlike in the binary case, it does *not* follow directly from a complementary slackness condition on the dual of the multiclass SVM. A key technical contribution that we provide to remedy this issue is a novel reparameterization of the SVM dual. The complete proof of Theorem 1 and this reparameterization is provided in Section 6.1.

We make a few additional remarks on the interpretation of Equation (14).

First, our proof shows a somewhat stronger conclusion: when Equation (12) holds, the multiclass SVM solutions $\hat{\mathbf{w}}_c, c \in [k]$ are same as the solutions to the following *simplex OvA-type classifier* (cf. Equation (8)):

$$\min_{\mathbf{w}_c} \ \frac{1}{2}\|\mathbf{w}_c\|_2^2 \qquad \text{sub. to} \quad \mathbf{x}_i^T \mathbf{w}_c \begin{cases} \geq \frac{k-1}{k} & , y_i = c, \\ \leq -\frac{1}{k} & , y_i \neq c, \end{cases} \quad \forall i \in [n], \tag{15}$$

for all $c \in [k]$. We note that the OvA-type classifier above can also be interpreted as a binary cost-sensitive SVM classifier [IMSV19] that enforces the margin corresponding to all other classes to be $(k-1)$ times smaller compared to the margin for the labeled class of the training data point. This simplex structure is illustrated in Figure 2, which evaluates the solution of the multiclass SVM on a 4-class Gaussian mixture model with isotropic noise covariance. The mean vectors are set to be mutually orthogonal and equal in norm, with SNR $\|\boldsymbol{\mu}\|_2 = 0.2\sqrt{p}$. We also set $n = 50$, $p = 1000$ to ensure sufficient effective overparameterization (in a sense that will be formally defined in subsequent sections). Figure 2 shows the inner product $\widehat{\mathbf{w}}_c^T \mathbf{x}$ drawn from 8 samples. These inner products are consistent with the simplex OvA structure defined in Equation (14), i.e. $\widehat{\mathbf{w}}_c^T \mathbf{x}_i = 3/4$ if $y_i = c$ and $\widehat{\mathbf{w}}_c^T \mathbf{x}_i = -1/4$ if $y_i \neq c$.

Second, Equation (14) shows that when Equation (12) holds, then the multiclass SVM solution $\mathbf{W}_{\mathrm{SVM}}$ has the same classification error as that of the minimum-norm interpolating solution. In other

words, we can show that the minimum-norm classifiers that interpolate the data with respect to *either* the one-hot representations $\mathbf{y}_i$ or the simplex representations $\widetilde{\mathbf{y}}_i$ of (10) have identical classification performance. This conclusion, stated as a corollary below, drives our classification error analysis in Section 4.

Corollary 1 (SVM=MNI)**.** *Under the same assumptions as in Theorem 1, and provided that the inequality in Equation (12) holds, it holds that* $\mathbb{P}_{e|c}(\mathbf{W}_{SVM}) = \mathbb{P}_{e|c}(\mathbf{W}_{MNI})$ *for all* $c \in [k]$*. Thus, the total classification errors of both solutions are equal:* $\mathbb{P}_e(\mathbf{W}_{SVM}) = \mathbb{P}_e(\mathbf{W}_{MNI})$*.*

The corollary follows directly by combining Theorem 1 with the following lemma applied with the choice $\alpha = 1, \beta = -1/k$. We include a detailed proof below for completeness.

Lemma 1. *For constants* $\alpha > 0, \beta$*, consider the MNI-solution* $\mathbf{w}_c^{\alpha,\beta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^+(\alpha\mathbf{v}_c + \beta\mathbf{1}), c \in [k]$ *corresponding to a target vector of labels* $\alpha\mathbf{v}_c + \beta\mathbf{1}_n$*. Let* $\mathbb{P}_{e|c}^{\alpha,\beta}, c \in [k]$ *be the class-conditional classification errors of the classifier* $\mathbf{w}^{\alpha,\beta}$*. Then, for any different set of constants* $\alpha' > 0, \beta'$*, it holds that* $\mathbb{P}_{e|c}^{\alpha,\beta} = \mathbb{P}_{e|c}^{\alpha',\beta'}, \forall c \in [k]$*.*

Proof. Note that $\mathbf{w}_c^{\alpha=1,\beta=0} = \mathbf{w}_{\mathrm{MNI},c}, c \in [k]$ and for arbitrary $\alpha > 0, \beta$, we have: $\mathbf{w}_c^{\alpha,\beta} = \alpha\mathbf{w}_{\mathrm{MNI},c} + \beta\mathbf{X}(\mathbf{X}^\top\mathbf{X})^+\mathbf{1}$. Moreover, it is not hard to check that $\mathbf{w}_{\mathrm{MNI},c}^\top\mathbf{x} \leq \max_{j\neq c}\mathbf{w}_{\mathrm{MNI},j}^\top\mathbf{x}$ if and only if $(\alpha\mathbf{w}_{\mathrm{MNI}c} + \mathbf{b})^\top\mathbf{x} \leq \max_{j\neq c}(\alpha\mathbf{w}_{\mathrm{MNI},j} + \mathbf{b})^\top\mathbf{x}$, for any $\mathbf{b} \in \mathbb{R}^p$. The claim then follows by choosing $\mathbf{b} = \beta\mathbf{X}(\mathbf{X}^\top\mathbf{X})^+\mathbf{1}$ and noting that $\alpha > 0, \beta$ were chosen arbitrarily. □

3.2 Connection to effective overparameterization

Theorem 1 establishes a *deterministic* condition that applies to any multiclass separable dataset as long as the data matrix $\mathbf{X}$ is full-rank. In this subsection, we show that the inequality in Equation (12) occurs with high-probability under both the GMM and MLM data models provided that there is sufficient effective overparameterization.

3.2.1 Gaussian mixture model

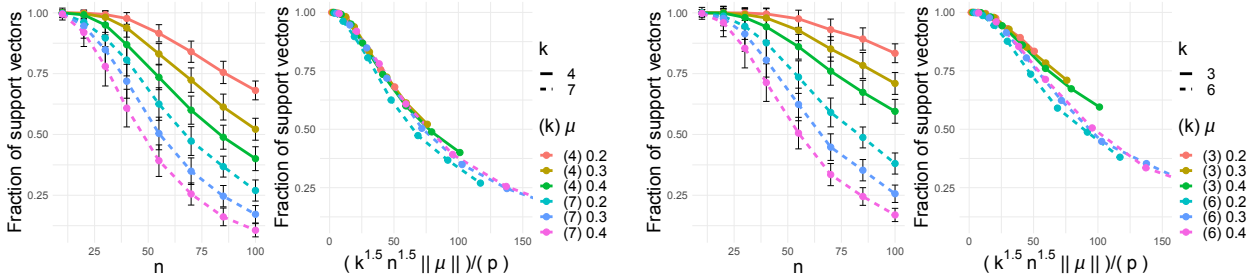
We assume a nearly equal-energy, equal-prior setting as detailed below.

Assumption 1 (Nearly equal energy/prior)**.** *We assume that the norms of the mean vectors are at the same order, i.e. for some large enough constants* $\{C_i\}_{i=1}^4$*, there exists a vector* $\boldsymbol{\mu}$ *such that the mean vectors satisfy* $(1 - \frac{1}{C_1})\|\boldsymbol{\mu}\|_2 \leq \|\boldsymbol{\mu}_c\|_2 \leq (1 + \frac{1}{C_2})\|\boldsymbol{\mu}\|_2, \forall c \in [k]$ *(equivalently, we have* $C_1 \leq \frac{\|\boldsymbol{\mu}_c\|_2}{\|\boldsymbol{\mu}_{c'}\|_2} \leq C_2$ *for all* $c, c' \in [k]$ *and large enough constants* $C_1, C_2 > 0$*). Moreover, the class priors are also at the same order, i.e. they satisfy* $(1 - \frac{1}{C_3})\frac{1}{k} \leq \pi_c \leq (1 + \frac{1}{C_4})\frac{1}{k}, \forall c \in [k]$ *(equivalently, we have* $C_3 \leq \frac{\pi_c}{\pi_{c'}} \leq C_4$ *for all* $c, c' \in [k]$ *and large enough constants* $C_3, C_4 > 0$*).*

Theorem 2. *Assume that the training set follows a multiclass GMM with* $\boldsymbol{\Sigma} = \mathbf{I}_p$*, Assumption 1 holds, and the number of training samples* n *is large enough. There exist constants* $c_1, c_2, c_3 > 1$ *and* $C_1, C_2 > 1$ *such that Equation (12) holds with probability at least* $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$*, provided that*

$$p > C_1 k^3 n \log(kn) + n - 1 \quad \text{and} \quad p > C_2 k^{1.5} n\sqrt{n}\|\boldsymbol{\mu}\|_2. \tag{16}$$

Theorem 2 establishes a set of two conditions under which Equation (12) and the conclusions of Theorem 1 hold, i.e. $\mathbf{W}_{\mathrm{SVM}} = \mathbf{W}_{\mathrm{MNI}}$. The first condition requires sufficient overparameterization $p = \Omega(k^3 n \log(kn))$, while the second one requires that the signal strength is not too large. Intuitively, we can understand these conditions as follows. Note that Equation (12) is satisfied provided that the inverse Gram matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ is "close" to identity, or any other positive-definite diagonal matrix. Recall from Equation (1) that $\mathbf{X} = \mathbf{M}\mathbf{Y} + \mathbf{Q} = \sum_{j=1}^k \boldsymbol{\mu}_j\mathbf{v}_j^T + \mathbf{Q}$ where $\mathbf{Q}$ is a $p \times n$ standard Gaussian matrix. The first inequality in Equation (16) (i.e. a lower bound on the data dimension p) is sufficient for $(\mathbf{Q}^T\mathbf{Q})^{-1}$ to have the desired property; the major technical challenge is that $(\mathbf{X}^T\mathbf{X})^{-1}$ involves additional terms that intricately depend on the label matrix $\mathbf{Y}$ itself. Our key technical contribution is showing that these extra terms do *not* drastically change the desired behavior, provided that the norms



(a) $k = 4$ and 7 (b) $k = 3$ and 6

Figure 3: Fraction of training examples satisfying Equation (14) (also called "support vectors") in the GMM case. The error bars show the standard deviation. Figure (a) considers $k = 4$ and 7, and Figure (b) considers $k = 3$ and 6. On the legend, "(4) 0.3" corresponds to $k = 4$ and $\|\boldsymbol{\mu}\|_2/\sqrt{p} = 0.2$. Observe that the curves nearly overlap when plotted versus $k^{1.5}n^{1.5}\|\boldsymbol{\mu}\|_2/p$ as predicted by the second condition in Equation (16) of Theorem 2.

of the mean vectors (i.e. signal strength) are sufficiently small. At a high-level we accomplish this with a recursive argument as follows. Denote $\mathbf{X}_0 = \mathbf{Q}$ and $\mathbf{X}_i = \sum_{j=1}^{i} \boldsymbol{\mu}_j \mathbf{v}_j^T + \mathbf{Q}$ for $i \in [k]$. Then, at each stage i of the recursion, we show how to bound quadratic forms involving $\left(\mathbf{X}_i^T \mathbf{X}_i\right)^{-1}$ using bounds established previously at stage $i-1$ on quadratic forms involving $\left(\mathbf{X}_{i-1}^T \mathbf{X}_{i-1}\right)^{-1}$. A critical property for the success of our proof strategy is the observation that the rows of $\mathbf{Y}$ are always orthogonal, that is, $\mathbf{v}_i^T \mathbf{v}_j = 0$, for $i \neq j$. The complete proof of the theorem is given in Section 6.2.

We first present numerical results that support the conclusions of Theorem 2. (In all our figures, we show averages over 100 Monte-Carlo realizations, and the error bars show the standard deviation at each point.) Figure 3(a) plots the fraction of support vectors satisfying Equation (14) as a function of training size n. We fix dimension $p = 1000$ and class priors $\pi = \frac{1}{k}$. To study how the outcome depends on the number of classes k and signal strength $\|\boldsymbol{\mu}\|_2$, we consider $k = 4, 7$ and three equal-energy scenarios where $\forall c \in [k] : \|\boldsymbol{\mu}_c\|_2 = \|\boldsymbol{\mu}\|_2 = \mu\sqrt{p}$ with $\mu = 0.2, 0.3, 0.4$. Observe that smaller μ results in larger proportion of support vectors for the same value of n. To verify our theorem's second condition (on the signal strength) in Equation (16), Figure 3(a) also plots the same set of curves over a re-scaled axis $k^{1.5}n^{1.5}\|\boldsymbol{\mu}\|_2/p$. The six curves corresponding to different settings nearly overlap in this new scaling, showing that the condition is order-wise tight. In Figure 3(b), we repeat the experiment in Figure 3(a) for different values of $k = 3$ and $k = 6$. Again, these curves nearly overlap when the x-axis is scaled according to the second condition on signal strength in Equation (16). We conjecture that our second condition on the signal strength is tight up to an extra $\sqrt{n}$ factor, which we believe is an artifact of the analysis[3]. We also believe that the k^3 factor in the first condition can be relaxed slightly to k^2 (as in the MLM case depicted in Figure 4, which considers a rescaled x-axis and shows *exact* overlap of the curves for all values of k). Sharpening these dependences on both k and n is an interesting direction for future work.

3.2.2 Multinomial logistic model

We now consider the MLM data model and anisotropic data covariance. Explicitly, the eigendecomposition of the covariance matrix is given by $\boldsymbol{\Sigma} = \sum_{i=1}^{p} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$, where $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_p]$. We also define the effective dimensions $d_2 := \|\boldsymbol{\lambda}\|_1^2/\|\boldsymbol{\lambda}\|_2^2$ and $d_\infty := \|\boldsymbol{\lambda}\|_1/\|\boldsymbol{\lambda}\|_\infty$. The following result contains sufficient conditions for the SVM and MNI solutions to coincide.

Theorem 3. *Assume n training samples following the MLM defined in (2). There exist constants c and $C_1, C_2 > 1$ such that Equation (12) holds with probability at least $(1 - \frac{c}{n})$ provided that*

$$d_\infty > C_1 k^2 n \log(kn) \text{ and } d_2 > C_2(\log(kn) + n). \tag{17}$$

[3]Support for this belief comes from the fact that [WT21] shows that $p > C_2\|\boldsymbol{\mu}\|_2 n$ is sufficient for the SVM = interpolation phenomenon to occur in the case of GMM and *binary* classification.

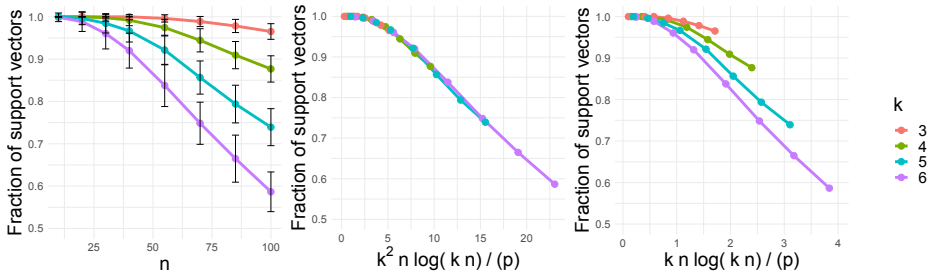


Figure 4: Fraction of training examples satisfying equality in the simplex label representation in Equation (14) in the MLM case with $\mathbf{\Sigma} = \mathbf{I}_p$. The middle plot shows that the curves overlap when plotted versus $k^2 n \log(kn)/p$ as predicted by Equation (18).

In fact, the only conditions we require on the generated labels is conditional independence.

For the isotropic case $\mathbf{\Sigma} = \mathbf{I}_p$, this implies that Equation (12) holds with probability at least $(1 - \frac{c}{n})$ provided that

$$p > C_1 k^2 n \log(kn). \tag{18}$$

The sufficient conditions in Theorem 3 require that the spectral structure in the covariance matrix $\mathbf{\Sigma}$ has sufficiently slowly decaying eigenvalues (corresponding to sufficiently large d_2), and that it is not too "spiky" (corresponding to sufficiently large d_∞). When $\mathbf{\Sigma} = \mathbf{I}_p$, the conditions reduce to sufficient overparameterization. For the special case of $k = 2$ classes, our conditions reduce to those in [HMX21] for binary classification. The dominant dependence on k, given by k^2, is a byproduct of the "unequal" margin in Equation (10). Figure 4 empirically verifies the sharpness of this factor.

The proof of Theorem 3 is provided in Appendix B. We now numerically validate our results in Theorem 3 in Figure 4, focusing on the isotropic case. We fix $p = 1000$, vary n from 10 to 100 and the numbers of classes from $k = 3$ to $k = 6$. We choose orthogonal mean vectors for each class with equal energy $\|\boldsymbol{\mu}\|_2^2 = p$. The left-most plot in Figure 4 shows the fraction of support vectors satisfying Equation (14) as a function of n. Clearly, smaller number of classes k results in higher proportion of support vectors with the desired property for the same number of measurements n. To verify the condition in Equation (18), the middle plot in Figure 4 plots the same curves over a re-scaled axis $k^2 n \log(kn)/p$ (as suggested by Equation (18)). We additionally draw the same curves over $kn \log(kn)/p$ in the right-most plot of Figure 3. Note the overlap of the curves in the middle plot. We now numerically validate our results in Theorem 3 in Figure 4, focusing on the isotropic case. We fix $p = 1000$, vary n from 10 to 100 and the numbers of classes from $k = 3$ to $k = 6$. We choose orthogonal mean vectors for each class with equal energy $\|\boldsymbol{\mu}\|_2^2 = p$. The left-most plot in Figure 4 shows the fraction of support vectors satisfying Equation (14) as a function of n. Clearly, smaller number of classes k results in higher proportion of support vectors with the desired property for the same number of measurements n. To verify the condition in Equation (18), the middle plot in Figure 4 plots the same curves over a re-scaled axis $k^2 n \log(kn)/p$ (as suggested by Equation (18)). We additionally draw the same curves over $kn \log(kn)/p$ in the right-most plot of Figure 3. Note the overlap of the curves in the middle plot.

3.3 Connection to Neural Collapse

In this section, we provide a distinct set of sufficient conditions on the feature vectors that guarantee Equation (12), and hence the conclusions of Theorem 1 hold. Interestingly, these sufficient conditions relate to the recently discovered, so called *neural-collapse* phenomenon that is empirically observed in the training process of overparameterized deep nets [PHD20] (see also e.g. [ZDZ+21, MPP20, HPD21, LS22, FHLS21a, FHLS21b, PL20b, GHNK21] for several recent follow-ups).

Corollary 2. *Recall the notation in Theorem 1. Assume exactly balanced data, that is $|\{i : y_i = c\}| = n/k$ for all $c \in [k]$. Also, assume that the following two conditions hold:*

- ***Feature collapse (NC1):** For each $c \in [k]$ and all $i \in [n] : y_i = c$, it holds that $\mathbf{x}_i = \boldsymbol{\mu}_c$, where $\boldsymbol{\mu}_c \triangleq \frac{k}{n} \sum_{i:y_i=c} \mathbf{x}_i$ is the "mean" vector of the corresponding class.*

- ***Simplex ETF structure (NC2)***: *The matrix of mean vectors* $\mathbf{M} := [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k]_{p\times k}$ *is the matrix of a simplex Equiangular Tight Frame (ETF), i.e. for some orthogonal matrix* $\mathbf{U}_{p\times k}$ *(with* $\mathbf{U}^T\mathbf{U} = \mathbf{I}_k$*) and* $\alpha \in \mathbb{R}$*, it holds that*

$$\mathbf{M} = \alpha\sqrt{\frac{k}{n}}\mathbf{U}\left(\mathbf{I}_k - \frac{1}{k}\mathbf{1}\mathbf{1}^T\right). \tag{19}$$

Then, the sufficient condition (12) *of Theorem 1 holds for the Gram matrix* $\mathbf{X}^T\mathbf{X}$.

Proof. For simplicity, denote the sample size of each class as $m := n/k$. Without loss of generality under the corollary's assumptions, let the columns of the feature matrix $\mathbf{X}$ be ordered such that $\mathbf{X} = [\mathbf{M}, \mathbf{M}, \ldots, \mathbf{M}] = \mathbf{M} \otimes \mathbf{1}_m^T$. Accordingly, we have $\mathbf{z}_c = (\mathbf{e}_c \otimes \mathbf{1}_m) - \frac{1}{k}(\mathbf{1}_k \otimes \mathbf{1}_m)$ where $\mathbf{e}_c$ is the c-th basis vector in $\mathbb{R}^k$. Then, the feature Gram matrix is computed as

$$\mathbf{X}^T\mathbf{X} = (\mathbf{M}^T\mathbf{M}) \otimes (\mathbf{1}_m\mathbf{1}_m^T) = \frac{\alpha^2}{m}\left(\mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T\right) \otimes (\mathbf{1}_m\mathbf{1}_m^T). \tag{20}$$

Observe here that we can write $\left(\mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T\right) = \mathbf{V}\mathbf{V}^T$ for $\mathbf{V} \in \mathbb{R}^{k\times(k-1)}$ having orthogonal columns (i.e. $\mathbf{V}^T\mathbf{V} = \mathbf{I}_{k-1}$) and $\mathbf{V}^T\mathbf{1}_k = \mathbf{0}_k$. Using this and the fact that $(\mathbf{V}\mathbf{V}^T)^+ = (\mathbf{V}\mathbf{V}^T)$, it can be checked from (20) that

$$(\mathbf{X}^T\mathbf{X})^+ = \frac{1}{\alpha^2 m}\left(\mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T\right) \otimes (\mathbf{1}_m\mathbf{1}_m^T). \tag{21}$$

Putting things together, we get, for any $c \in [k]$, that

$$(\mathbf{X}^T\mathbf{X})^+\mathbf{z}_c = \frac{1}{\alpha^2 m}\left((\mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T) \otimes (\mathbf{1}_m\mathbf{1}_m^T)\right)(\mathbf{e}_c \otimes \mathbf{1}_m) = \frac{1}{\alpha^2}\left(\mathbf{e}_c - \frac{1}{k}\mathbf{1}_k\right) \otimes \mathbf{1}_m = \frac{1}{\alpha^2}\mathbf{z}_c.$$

Therefore, it follows immediately that

$$\mathbf{z}_c \odot \mathbf{M}^+\mathbf{z}_c = \frac{1}{\alpha^2}\mathbf{z}_c \odot \mathbf{z}_c > \mathbf{0},$$

as desired. This completes the proof. □

It might initially appear that the structure of the feature vectors imposed by the properties NC1 and NC2 is too specific to be relevant in practice. To the contrary, [PHD20] showed via a principled experimental study that these properties occur at the last layer of overparameterized deep nets across several different data sets and DNN architectures. Specifically, the experiments conducted in [PHD20] suggest that training overparameterized deep nets on classification tasks with CE loss in the absence of weight decay (i.e. without explicit regularization) results in learned feature representations in the final layer that converge[4] to the ETF structure described by NC1 and NC2. Furthermore, it was recently shown in [GHNK21] that the neural collapse phenomenon continues to occur when the last-layer features of a deep net are trained with the recently proposed *supervised contrastive loss* (SCL) function [KTW+20] and a linear model is independently trained on these learned last-layer features. (In fact, [GHNK21, KTW+20] showed that this self-supervised procedure can yield superior generalization performance compared to CE loss.)

To interpret Corollary 2 in view of these findings, consider the following two-stage classification training process:

- First, train (without weight-decay and continuing training beyond the interpolation regime) the last-layer feature representations of an overparameterized deep-net with either CE or SCL losses.
- Second, taking as inputs those learned feature representations of the first stage, train a linear multiclass classifier (often called the "head" of the deep-net) with CE loss.

[4]Here, "convergence" is with respect to an increasing number of training epochs. Since the architecture is overparameterized, it can perfectly separate the data. Hence, the training 0-1 error can be driven to zero. Nevertheless, training continues despite having achieved zero 0-1 training error, since the CE loss continues to drop. [PHD20] refers to this regime as the terminal phase of training (TPT). In sum, [PHD20] show that neural collapse is observed in TPT.

Then, from Corollary 2, the resulting classifier from this two-stage process interpolates the simplex label representation, and the classification accuracy is the same as if we had used the square loss in the second stage of the above training process. Thus, our results lend strong theoretical justification to the empirical observation that *square-loss and CE loss yield near-identical performance in large-scale classification tasks [Rif02, RK04, HB20, PL20a]*.

4 Generalization bounds

In this section, we derive non-asymptotic bounds on the error of the MNI classifier for data generated from both GMM and MLM, as well as a natural setting in which the class means follow the simplex-ETF geometry.

4.1 Gaussian mixture model

We present classification error bounds under the additional assumption of mutually incoherent means.

Assumption 2 (Mutually incoherent means)**.** *Let* $M = \max_{i \neq j} \frac{|\boldsymbol{\mu}_i^T \boldsymbol{\mu}_j|}{\|\boldsymbol{\mu}_i\|_2 \|\boldsymbol{\mu}_j\|_2}$ *be the mutual coherence of mean vectors. Then, we assume that there exists a large absolute constant* $C > 0$ *such that* $M \leq 1/C$.

We remark that mutual incoherence assumptions have appeared in a completely different context, i.e. across feature vectors, in the compressive sensing literature (e.g. for sparse signal recovery) [DET05, Tro06]. There, the number of feature vectors is typically greater than the dimension of each feature vector and so the mutual incoherence suffers from fundamental lower bounds [Wel74]. In our setting, the incoherence assumption applies to the class-mean vectors. Note that the number of mean vectors (k) is always smaller than the dimension of each vector (p) and so Welch's lower bound does not apply, making our assumption reasonable.

Theorem 4. *Let Assumptions 1 and 2, as well as the condition in Equation (16) hold. Further assume constants* $C_1, C_2, C_3 > 1$ *such that* $\left(1 - \frac{C_1}{\sqrt{n}} - \frac{C_2 n}{p}\right)\|\boldsymbol{\mu}\|_2 > C_3 \min\{\sqrt{k}, \sqrt{\log(2n)}\}$. *Then, there exist additional constants* c_1, c_2, c_3 *and* $C_4 > 1$ *such that both the MNI solution* $\mathbf{W}_{MNI}$ *and the multiclass SVM solution* $\mathbf{W}_{SVM}$ *satisfy*

$$\mathbb{P}_{e|c} \leq (k-1) \exp\left(-\|\boldsymbol{\mu}\|_2^2 \frac{\left(\left(1 - \frac{C_1}{\sqrt{n}} - \frac{C_2 n}{p}\right)\|\boldsymbol{\mu}\|_2 - C_3 \min\{\sqrt{k}, \sqrt{\log(2n)}\}\right)^2}{C_4 \left(1 + \frac{kp}{n\|\boldsymbol{\mu}\|_2^2}\right)} \right) \tag{22}$$

with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$, *for every* $c \in [k]$. *Moreover, the same bound holds for the total classification error* $\mathbb{P}_e$.

For large enough n, Theorem 4 reduces to the results in [WT21] when $k = 2$ (with slightly different constants). There are two major challenges in the proof of Theorem 4, which is presented in Appendix C.1. First, in contrast to the binary case the classification error does *not* simply reduce to bounding correlations between vector means $\boldsymbol{\mu}_c$ and their estimators $\hat{\mathbf{w}}_c$. Second, just as in the proof of Theorem 2, technical complications arise from the multiple mean components in the training data matrix $\mathbf{X}$. We use a variant of the recursion-based argument described in Section 6.2 to obtain our final bound.

4.1.1 A possible extension to anisotropic noise covariances

Up to this point, we have concentrated on GMM data with isotropic noise, i.e. the noise covariance matrix in Equation (1) is such that $\boldsymbol{\Sigma} = \mathbf{I}_p$. It is crucial to note that, even in the case of isotropic noise, the entire data covariance matrix $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$ for GMM data is anisotropic, as it exhibits spikes in the direction of the mean vectors. Thus, it already models highly correlated features. This already makes the analyses challenging both at the level of establishing equivalence of SVM to MNI as well as deriving generalization bounds for the MNI (analogous to the challenges faced in the initial analyses of benign

overfitting for regression [BLLT20, HMRT19]). Based on this, we now make a brief comment on the possibility of extending Theorem 4 to anisotropic GMM data. Although a comprehensive examination is beyond the scope of this paper, we provide evidence that our analysis can serve as a foundation for such extensions.

As a starting point, we note that the necessary and sufficient equivalence conditions of Theorem 1 still hold (as they are deterministic and require no assumptions on the data). We sketch here a possible proof argument to work from Theorem 1 and prove a variant of Theorem 2 for anisotropic noise covariance. Let $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ be the covariance eigen-decomposition, where $\mathbf{V}$ is orthogonal and $\mathbf{\Lambda}$ is diagonal with entries given by the eigenvalues $\{\lambda_j\}_{j=1}^p$. With this, we can project the mean vectors $\boldsymbol{\mu}_c$ of the GMM to the space spanned by the eigenvector basis $\mathbf{v}_j, j \in [p]$ (aka columns of $\mathbf{V}$). Concretely, we can express $\boldsymbol{\mu}_c$ as $\sum_{j=1}^p \beta_j \mathbf{v}_j$. Then, we may use this decomposition to prove a variant of Lemma 2. Recall that to prove the higher-order terms in Lemma 2, we need to start from deriving bounds for 0-order terms in Lemma 7. When $\mathbf{\Sigma}$ is anisotropic, the bounds in Lemma 7 will have two main changes. First, the bounds will involve *signal strength in the direction of* $\mathbf{\Sigma}$ defined as $\sum_{j=1}^p \lambda_j \beta_j^2$. This is the result of projecting the mean vectors to the space spanned by the eigenvectors of $\mathbf{\Sigma}$. Second, the bounds will include effective ranks, e.g. $r_k := (\sum_{i>k}^p \lambda_i)/\lambda_{k+1}$ and $R_k := (\sum_{i>k}^p \lambda_i)^2/(\sum_{i>k}^p \lambda_i^2)$. Effective ranks play important role in benign overfitting and the equivalence between SVM and MNI [BLLT20, MNS+21]. Lemma 4 in [WT21] provides bounds for the 0-order terms in Lemma 7 under anisotropic covariance. We show one examples here to see the adjustment. The upper bound for $t_{jj}^{(0)}$ changes from $\frac{C_1 n \|\boldsymbol{\mu}\|_2^2}{p}$ to $\frac{C_2 n \sum_{j=1}^p \lambda_j \beta_j^2}{\|\lambda\|_1}$, where λ is the vector with λ_i as entries. Note that $\frac{n \sum_{j=1}^p \lambda_j \beta_j^2}{\|\lambda\|_1}$ becomes $\frac{n\|\boldsymbol{\mu}\|_2^2}{p}$ when $\mathbf{\Sigma} = \mathbf{I}_p$. Similar changes apply to other terms in Lemma 7. The 0-order bounds in Lemma 7 can then be used to derive higher-order bounds in Lemma 2. Similar to the binary results in [MNS+21, WT21], the equivalence between MNI and SVM requires large effective ranks and benign overfitting requires large signal strength in the direction of $\mathbf{\Sigma}$. However, a detailed analysis of this general setting is beyond the scope of this paper.

4.2 Multinomial logistic model

In this section, we present our error analysis of the MNI classifier when data is generated by the MLM. Importantly, for this case we consider more general anisotropic structure in the covariance matrix $\mathbf{\Sigma} := \boldsymbol{U}\mathbf{\Lambda}\boldsymbol{U}^\top$. We begin by carrying over the assumptions made from the binary-case analysis in [MNS+21], beginning with a natural assumption of s-sparsity.

Assumption 3 (s-sparse class means)**.** *We assume that all of the class means $\boldsymbol{\mu}_c, c \in [k]$ are s-sparse in the basis given by the eigenvectors of $\mathbf{\Sigma}$. In other words, we have*

$$\boldsymbol{U}^{-1}\boldsymbol{\mu}_{c,j} = 0 \text{ if } j > s.$$

This s-sparse assumption is also made in corresponding works on regression (e.g. for the results for the anisotropic case in [HMRT19]) and shown to be necessary in an approximate sense for consistency of MSE of the minimum-ℓ_2-norm interpolation arising from bias [TB20]. Next, we make a special assumption of bi-level structure in the covariance matrix.

Assumption 4 (Bi-level ensemble)**.** *We assume that the eigenvalues of the covariance matrix, given by $\boldsymbol{\lambda}$, have a bilevel structure. In particular, our bi-level ensemble is parameterized by (n, m, q, r) where $m > 1$, $0 \le r < 1$ and $0 < q < (m - r)$. We set parameters $p = n^m$, $s = n^r$ and $a = n^{-q}$. Then, the eigenvalues of the covariance matrix are given by*

$$\lambda_j = \begin{cases} \lambda_H := \frac{ap}{s}, & 1 \le j \le s \\ \lambda_L := \frac{(1-a)p}{p-s}, & \text{otherwise.} \end{cases}$$

We will fix (m, q, r) and study the classification error as a function of n. While the bi-level ensemble structure is not in principle needed for complete statements of results, it admits particularly clean characterizations of classification error rates as well as easily interpretable conditions for consistency[5]*.*

[5] See [MNS+21] for additional context on the bi-level ensemble and examples of its manifestation in high-dimensional machine learning models.

Assumption 4 splits the covariance spectrum in a small set of large eigenvalues λ_H and the remaining large set of small eigenvalues λ_L. The bi-level ensemble is friendly to consistency of the MNI solution for three reasons: a) the number of small eigenvalues is much larger than the sample size, b) the ratio between the large-valued and small-valued eigenvalues grows with the sample size n, and c) the number of large-valued eigenvalues is exponentially small relative to the sample size n. Note that condition a) facilitates benign overfitting of noise (as first pointed out in the more general anisotropic case by [BLLT20]), while conditions b) and c) facilitate signal recovery. To verify these conditions more quantitatively, note that: a) the number of small eigenvalues is on the order of $p \gg n$, b) the ratio between the large-valued and small-valued eigenvalues can be verified to be on the order of n^{m-q-r} which grows with n, and c) the number of large-valued eigenvalues is equal to $s = n^r$, which is exponentially smaller than n.

Finally, we imbue the above assumptions with an equal energy and orthogonality assumption, as in the GMM case. These assumptions are specific to the multiclass task, and effectively subsume Assumption 3.

Assumption 5 (Equal energy and orthogonality)**.** *We assume that the class means are equal energy, i.e.* $\|\boldsymbol{\mu}\|_2 = 1/\sqrt{\lambda_H}$ *for all* $c \in [k]$*, and are orthogonal, i.e.* $\boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j = 0$ *for all* $i \neq j \in [k]$*. Together with Assumptions 3 and 4, a simple coordinate transformation gives us*

$$\boldsymbol{\mu}_c = \frac{1}{\sqrt{\lambda_H}} \boldsymbol{e}_{j_c} \;\; \text{for some } j_c \in [s],\; j_c \neq j_{c'} \text{ for all } c \neq c' \in [k], \text{ and}$$
$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}$$

without loss of generality. The normalization by the factor $\frac{1}{\sqrt{\lambda_H}}$ *is done to ensure that the signal strength is equal to* 1*, i.e.* $\mathbb{E}[(\mathbf{x}^\top \boldsymbol{\mu}_c)^2] = 1$ *for all* $c \in [k]$*.*

Under these assumptions, we state our main result for the total classification error of MLM. Our error bounds will be on the *excess* risk over and above the Bayes error rate incurred by the optimal classifier $\{\widehat{\mathbf{w}}_c = \boldsymbol{\mu}_c\}_{c\in[k]}$, which we denote by $\mathbb{P}_{e,\mathsf{Bayes}}$.

Theorem 5. *Under Assumptions 4 and 5, there is a universal constant* c_k *(that may depend on* k*, but not* n *or* p*) such that the total excess classification error of* $\mathbf{W}_{MNI}$ *and* $\mathbf{W}_{SVM}$ *under the MLM model is given by*

$$\mathbb{P}_e - \mathbb{P}_{e,\mathsf{Bayes}} \leq k^2 \left(\frac{1}{2} - \frac{1}{\pi} \mathsf{tan}^{-1}(\mathsf{SNR}(n)) \right), \quad \textit{where} \tag{23}$$
$$\mathsf{SNR}(n) \geq c_k (\log n)^{-1/2} \cdot n^{\frac{\min\{(m-1),(2q+r-1),(2q+2r-3/2)\}}{2} + (1-r) - q}, \; q > (1-r)$$

for $q > 1 - r$*.*

The proof of Theorem 5 is presented in Section 6.3. We will show in the subsequent Section 5 that, although the rate in Equation (23) is worse in its dependence on q and r than for the equivalent binary classification problem, the conditions for benign overfitting turn out to coincide in the regime where we keep k constant with respect to n.

4.3 Means following the simplex-ETF geometry

Next, we derive generalization bounds under an entirely different assumption on the geometry of mean vectors. Specifically, we consider the setting in which the mean vectors follow the simplex ETF geometry structure that was discussed in Section 3.3. Recall, this setting is particularly interesting and relevant to practice, as the ETF geometry describes the geometry of learnt class-mean embeddings of deep-nets when trained with the CE loss to completion (i.e., beyond achieving zero $0-1$ training error) [PHD20].

Theorem 6. *Let the nearly equal energy/prior Assumption 1 and the conditions in Equation (16) hold. Additionally, assume the means form a ETF structure, i.e.* $\boldsymbol{\mu}_i^T \boldsymbol{\mu}_i = -(k-1)\boldsymbol{\mu}_i^T \boldsymbol{\mu}_j$*, for* $i \neq j$*. Further assume constants* $C_1, C_2, C_3 > 1$ *such that* $\left(1 - \frac{C_1}{\sqrt{n}} - \frac{C_2 n}{p}\right)\|\boldsymbol{\mu}\|_2 > C_3 \min\{\sqrt{k}, \sqrt{\log(2n)}\}$*. Then, there*

exist additional constants c_1, c_2, c_3 *and* $C_4 > 1$ *such that both the MNI solution* $\mathbf{W}_{MNI}$ *and the multiclass SVM solution* $\mathbf{W}_{SVM}$ *satisfy*

$$\mathbb{P}_{e|c} \leq (k-1)\exp\left(-\|\boldsymbol{\mu}\|_2^2 \frac{\left(\left(1-\frac{C_1}{\sqrt{n}}-\frac{C_2 n}{p}\right)\|\boldsymbol{\mu}\|_2 - C_3\min\{\sqrt{k},\sqrt{\log(2n)}\}\right)^2}{C_4\left(1+\frac{kp}{n\|\boldsymbol{\mu}\|_2^2}\right)}\right) \tag{24}$$

with probability at least $1-\frac{c_1}{n}-c_2 k e^{-\frac{n}{c_3 k^2}}$*, for every* $c \in [k]$*. Moreover, the same bound holds for the total classification error* $\mathbb{P}_e$*.*

The proof of this theorem is provided in Appendix C.2. The non-zero inner products between $\boldsymbol{\mu}_i^T$ and $\boldsymbol{\mu}_j$ contribute "negatively" to the signal $\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i$. This negative contribution can be negated because of the simplex ETF structure $\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i = -(k-1)\boldsymbol{\mu}_i^T\boldsymbol{\mu}_j$, hence the bounds in Theorem 4 still hold.

5 Conditions for benign overfitting

Thus far, we have studied the classification error of the MNI classifier under the GMM data model (Theorem 4), and shown equivalence of the multiclass SVM and MNI solutions (Theorems 1, 2 and Corollary 1). Combining these results, we now provide sufficient conditions under which the classification error of the multiclass SVM solution (also of the MNI) approaches 0 as the number of parameters p increases. First, we state our sufficient conditions for harmless interpolation under the GMM model — these arise as a consequence of Theorem 4, and the proof is provided in Appendix C.3.

Corollary 3. *Let the same assumptions as in Theorem 4 hold. Then, for finite number of classes* k *and sufficiently large sample size* n*, there exist positive constants* c_i*'s and* C_i*'s* > 1*, such that the multiclass SVM classifier* $\mathbf{W}_{SVM}$ *in (7) satisfies the simplex interpolation constraint in (14) and its total classification error approaches* 0 *as* $\left(\frac{p}{n}\right) \to \infty$ *with probability at least* $1-\frac{c_1}{n}-c_2 k e^{-\frac{n}{c_3 k^2}}$*, provided that the following conditions hold:*
(1). When $\|\boldsymbol{\mu}\|_2^2 > \frac{kp}{n}$*,*

$$\frac{n}{C_1 k}\|\boldsymbol{\mu}\|_2^2 > p > \max\{C_2 k^3 n\log(kn) + n - 1, C_3 k^{1.5} n^{1.5}\|\boldsymbol{\mu}\|_2\}.$$

(2). When $\|\boldsymbol{\mu}\|_2^2 \leq \frac{kp}{n}$*,*

$$p > \max\{C_2 k^3 n\log(kn) + n - 1, C_3 k^{1.5} n^{1.5}\|\boldsymbol{\mu}\|_2, \frac{n\|\boldsymbol{\mu}\|_2^2}{k}\},$$
$$\text{and } \|\boldsymbol{\mu}\|_2^4 \geq C_4\left(\frac{p}{n}\right)^{\alpha}, \quad \text{for } \alpha > 1.$$

When n *is fixed, the conditions for benign overfitting for* $\mathbf{W}_{SVM}$ *become* $\|\boldsymbol{\mu}\|_2 = \Theta(p^{\beta})$ *for* $\beta \in (1/4, 1)$*.*

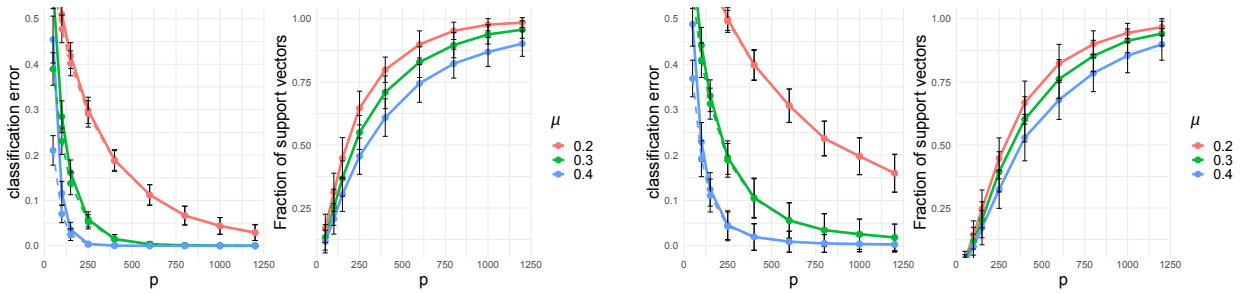
Note that the upper bound on $\|\boldsymbol{\mu}\|_2$ comes from the conditions that make SVM=MNI in Theorem 2; indeed, a distinct corollary of Theorem 4 is that $\mathbf{W}_{\text{MNI}}$ overfits benignly with sufficient signal strength $\|\boldsymbol{\mu}\|_2 = \Omega(p^{1/4})$. We can compare our result with the binary case [WT21]. When k and n are both finite, the condition $\|\boldsymbol{\mu}\|_2 = \Theta(p^{\beta})$ for $\beta \in (1/4, 1)$ is the same as the binary result.

Next, we state our sufficient and necessary conditions for harmless interpolation under the MLM model.

Corollary 4. *Let the same assumptions as in Theorem 5 hold. Then, for finite number of classes* k*, the following parameters of the bilevel ensemble (Assumption 4) ensure that the total classification error of* $\mathbf{W}_{SVM}$ *approaches* 0 *as* $n \to \infty$*:*

$$p > 1 \text{ and } q < (1-r) + \frac{(m-1)}{2}. \tag{25}$$

Further, when $q > (1-r)$*, the same conclusion holds for* $\mathbf{W}_{MNI}$*.*



(a) $k = 4$ (b) $k = 6$

Figure 5: Evolution of total classification error and fraction of support vectors as a function of p in the GMM case. Figure (a) considers $k = 4$ and Figure (b) considers $k = 6$. We consider the energy of all class means to be $\|\boldsymbol{\mu}\|_2 = \mu\sqrt{p}$, where $\mu = 0.2, 0.3$ and 0.4. Observe that the total classification error approaches 0 and the fraction of support vectors approaches 1 as p gets larger.

Proof. We work from Equation (23) of Theorem 5. For $\mathbb{P}_e - \mathbb{P}_{e,\mathsf{Bayes}} \to 0$ as $n \to \infty$, we require the exponent $\frac{\min\{(m-1),(2q+r-1),(2q+2r-3/2)\}}{2} + (1-r) - q \geq 0$. If $2q + 2r - 3/2$ is the minimizer, we would have $q + r - 3/4 + 1 - r - q = 1/4$, in which case the inequality is satisfied. If $2q + r - 1$ is the minimizer, we would have $q + r/2 - 1/2 + 1 - r - q = \frac{1-r}{2} > 0$, in which case the inequality is again satisfied. Otherwise, we have $\frac{m-1}{2} + 1 - r - q > 0$, which implies $q < (1-r) + \frac{(m-1)}{2}$. □

We can again compare our result with the binary case [MNS+21]: when k is finite, the conditions in Equation (25) are identical to those for the binary case. We also note that while Theorem 5 only provides an upper bound on MLM classification error, [MNS+21] provides lower bounds for the binary case that automatically apply to the MLM for the special case $k = 2$. While there is a gap between the non-asymptotic rates, the necessary conditions for consistency coincide with Equation (25). Therefore, Equation (25) encapsulates sufficient and necessary conditions for consistency when k is kept constant with respect to n. Moreover, as [MNS+21] show, the condition $q \leq (1-r)$ would be requirement for a corresponding *regression* task to generalize; consequently, Corollary 4 shows that *multiclass classification can generalize even when regression does not.*

We particularly note that, Corollaries 3 and 4 imply benign overfitting in regimes that cannot be explained by classical *training-data-dependent* bounds based on the margin [SFBL98]. While the shortcomings of such margin-based bounds in the highly overparameterized regime are well-documented, e.g. [DR17], we provide a brief description here for completeness. For the MLM, [MNS+21, Section 6] shows (for the binary case) that margin-based bounds could only predict harmless interpolation if we had the significantly stronger condition $q \leq (1-r)$ (also required for consistency of the corresponding regression task). For the GMM, we verify here that the margin-based bounds could only predict benign overfitting if we had the significantly stronger condition $\beta \in (1/2, 1)$ (see also [WT21, Section 9.1]): in the regime where SVM = MNI, the margin is exactly equal to 1. The margin-based bounds (as given in, e.g. [BM03]), can be verified to scale as $\mathcal{O}\left(\sqrt{\frac{\mathrm{trace}(\boldsymbol{\Sigma}_{\mathrm{un}})}{n\|\boldsymbol{\Sigma}_{\mathrm{un}}\|_2}}\right)$ with high probability, where $\boldsymbol{\Sigma}_{\mathrm{un}} := \mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right]$ denotes the *unconditional* covariance matrix under the GMM. In the case of the binary GMM and isotropic noise covariance, an elementary calculation shows that the spectrum of $\boldsymbol{\Sigma}_{\mathrm{un}}$ is given by $\left[\|\boldsymbol{\mu}\|_2^2 + 1 \quad 1 \quad \ldots \quad 1\right]$; plugging this into the above bound requires $\|\boldsymbol{\mu}\|_2^2 \gg \frac{p}{n}$ for the margin-based upper bound to scale as $o(1)$. This clearly does not explain benign overfitting when SVM = MNI, which we showed requires $\|\boldsymbol{\mu}\|_2^2 \leq \frac{p}{n}$.

Finally, we present numerical illustrations validating our benign overfitting results in Corollary 3. In Figure 5(a), we set the number of classes $k = 4$. To guarantee sufficient overparameterization, we fix $n = 40$ and vary p from 50 to 1200. We simulate 3 different settings for the mean matrices: each has orthogonal and equal-norm mean vectors $\|\boldsymbol{\mu}\|_2 = \mu\sqrt{p}$, with $\mu = 0.2, 0.3$ and 0.4. Figure 5 plots the classification error as a function of p for both MNI estimates (solid lines) and multiclass SVM solutions (dashed lines). Different colors correspond to different mean norms. The solid and dashed

curves almost overlap as predicted from our results in Section 3. We verify that as p increases, the classification error decreases towards zero. Observe that the fraction of support vectors approaches 1 as p gets larger. Further, the classification error goes to zero very fast when μ is large, but then the proportion of support vectors increases at a slow rate. In contrast, when μ is small, the proportion of support vectors increases fast, but the classification error decreases slowly. Figure 5(b) uses the same setting as in Figure 5(a) except for setting $k = 6$ and $n = 30$. Observe that the classification error continues to go to zero and the proportion of support vectors continues to increase, but both become slower as the number of classes is now greater.

6 Proofs of main results

In this section, we provide the proofs of Theorems 1, 2 and 5. The proof techniques we developed for these results convey novel technical ideas that also form the core of the rest of the proofs, which we defer to the Appendix.

6.1 Proof of Theorem 1

Argument sketch. We split the proof of the theorem in three steps. To better convey the main ideas, we first outline the three steps in this paragraph before discussing their details in the remaining of this section.

Step 1: The first key step to prove Theorem 1 is constructing a new parameterization of the dual of the multiclass SVM, which we show takes the following form:

$$\begin{aligned} \max_{\boldsymbol{\beta}_c \in \mathbb{R}^n, c \in [k]} \quad & \sum_{c \in [k]} \boldsymbol{\beta}_c^T \mathbf{z}_c - \frac{1}{2} \|\mathbf{X} \boldsymbol{\beta}_c\|_2^2 \\ \text{sub. to} \quad & \beta_{y_i,i} = -\sum_{c \neq y_i} \beta_{c,i}, \ \forall i \in [n] \quad \text{and} \quad \boldsymbol{\beta}_c \odot \mathbf{z}_c \geq \mathbf{0}, \forall c \in [k]. \end{aligned} \tag{26}$$

Here, for each $c \in [k]$ we let $\boldsymbol{\beta}_c = [\beta_{c,1}, \beta_{c,2}, \ldots, \beta_{c,n}] \in \mathbb{R}^n$. We also show by complementary slackness the following implication for any *optimal* $\beta^*_{c,i}$ in (26):

$$z_{c,i} \beta^*_{c,i} > 0 \implies (\hat{\mathbf{w}}_{y_i} - \hat{\mathbf{w}}_c)^T \mathbf{x}_i = 1. \tag{27}$$

Thus, to prove Equation (13), it will suffice showing that $z_{c,i} \beta^*_{c,i} > 0, \forall i \in [n], c \in [k]$ provided that Equation (12) holds.

Step 2: To do this, we prove that the *unconstrained* maximizer in (26), that is $\hat{\boldsymbol{\beta}}_c = (\mathbf{X}^T \mathbf{X})^+ \mathbf{z}_c$, $\forall c \in [k]$ is feasible, and therefore optimal, in (26). Now, note that Equation (12) is equivalent to $\mathbf{z}_c \odot \hat{\boldsymbol{\beta}}_c > 0$; thus, we have found that $\hat{\boldsymbol{\beta}}_c, c \in [k]$ further satisfies the n *strict* inequality constraints in (27) which completes the proof of the first part of the theorem (Equation (13)).

Step 3: Next, we outline the proof of Equation (14). We consider the simplex-type OvA-classifier in (15). The proof has two steps. First, using similar arguments to what was done above, we show that when Equation (12) holds, then all the inequality constraints in (15) are active at the optimal. That is, the minimizers $\mathbf{w}_{\text{OvA},c}$ of (15) satisfy Equation (14). Second, to prove that Equation (14) is satisfied by the minimizers $\hat{\mathbf{w}}_c$ of the multiclass SVM in (7), we need to show that $\mathbf{w}_{\text{OvA},c} = \hat{\mathbf{w}}_c$ for all $c \in [k]$. We do this by showing that, under Equation (12), the duals of (7) and (15) are equivalent. By strong duality, the optimal costs of the primal problems are also the same. Then, because a) the objective is the same for the two primals, b) $\mathbf{w}_{\text{OvA},c}$ is feasible in (15) and c) (7) is strongly convex, we can conclude with the desired.

Step 1: Key alternative parameterization of the dual. We start by writing the dual of the multiclass SVM, repeated here for convenience:

$$\min_{\mathbf{W}} \frac{1}{2} \|\mathbf{W}\|_F^2 \quad \text{sub. to} \quad (\mathbf{w}_{y_i} - \mathbf{w}_c)^\top \mathbf{x}_i \geq 1, \ \forall i \in [n], c \in [k] : c \neq y_i. \tag{28}$$

We have dual variables $\{\lambda_{c,i}\}$ for every $i \in [n], c \in [k] : c \neq y_i$ corresponding to the constraints on the primal form above. Then, the dual of the multiclass SVM takes the form

$$\max_{\lambda_{c,i} \geq 0} \sum_{i \in [n]} \Big(\sum_{\substack{c \in [k] \\ c \neq y_i}} \lambda_{c,i} \Big) - \frac{1}{2} \sum_{c \in [k]} \Big\| \sum_{i \in [n] : y_i = c} \Big(\sum_{\substack{c' \in [k] \\ c' \neq y_i}} \lambda_{c',i} \Big) \mathbf{x}_i - \sum_{i \in [n] : y_i \neq c} \lambda_{c,i} \mathbf{x}_i \Big\|_2^2. \tag{29}$$

Let $\hat{\lambda}_{c,i}, i \in [n], c \in [k] : c \neq y_i$ be maximizers in Equation (29). By complementary slackness, we have

$$\hat{\lambda}_{c,i} > 0 \implies (\hat{\mathbf{w}}_{y_i} - \hat{\mathbf{w}}_c)^\top \mathbf{x}_i = 1. \tag{30}$$

Thus, it will suffice to prove that $\hat{\lambda}_{c,i} > 0, \forall i \in [n], c \in [k] : c \neq y_i$ provided that (12) holds.

It is challenging to work directly with Equation (29) because the variables $\lambda_{c,i}$ are coupled in the objective function. Our main idea is to re-parameterize the dual objective in terms of new variables $\{\beta_{c,i}\}$, which we define as follows for all $c \in [k]$ and $i \in [n]$:

$$\beta_{c,i} = \begin{cases} \sum_{c' \neq y_i} \lambda_{c',i} & , y_i = c, \\ -\lambda_{c,i} & , y_i \neq c. \end{cases} \tag{31}$$

For each $c \in [k]$, we denote $\boldsymbol{\beta}_c = [\beta_{c,1}, \beta_{c,2}, \ldots, \beta_{c,n}] \in \mathbb{R}^n$. With these, we show that the dual objective becomes

$$\sum_{c \in [k]} \boldsymbol{\beta}_c^\top \mathbf{z}_c - \frac{1}{2} \sum_{c \in [k]} \Big\| \sum_{i \in [n]} \beta_{c,i} \mathbf{x}_i \Big\|_2^2 = \sum_{c \in [k]} \boldsymbol{\beta}_c^\top \mathbf{z}_c - \frac{1}{2} \|\mathbf{X} \boldsymbol{\beta}_c\|_2^2. \tag{32}$$

The equivalence of the quadratic term in $\boldsymbol{\beta}$ is straightforward. To show the equivalence of the linear term in $\boldsymbol{\beta}$, we denote $A := \sum_{i \in [n]} \Big(\sum_{c \in [k], c \neq y_i} \lambda_{c,i} \Big)$, and simultaneously get

$$A = \sum_{i \in [n]} \beta_{y_i,i} \qquad \text{and} \qquad A = \sum_{i \in [n]} \sum_{c \neq y_i} (-\beta_{c,i}),$$

by the definition of variables $\{\beta_{c,i}\}$ in Equation (31). Then, we have

$$\begin{aligned} A = \frac{k-1}{k} \cdot A + \frac{1}{k} \cdot A &= \frac{k-1}{k} \sum_{i \in [n]} \beta_{y_i,i} + \frac{1}{k} \sum_{i \in [n]} \sum_{c \neq y_i} (-\beta_{c,i}) \\ &\overset{\text{(i)}}{=} \sum_{i \in [n]} \mathbf{z}_{y_i,i} \beta_{y_i,i} + \sum_{i \in [n]} \sum_{c \neq y_i} \mathbf{z}_{c,i} \beta_{c,i} \\ &= \sum_{i \in [n]} \sum_{c \in [k]} \mathbf{z}_{c,i} \beta_{c,i} = \sum_{c \in [k]} \boldsymbol{\beta}_c^\top \mathbf{z}_c. \end{aligned}$$

Above, inequality (i) follows from the definition of $\mathbf{z}_c$ in Equation (11), rewritten coordinate-wise as:

$$z_{c,i} = \begin{cases} \frac{k-1}{k}, & y_i = c, \\ -\frac{1}{k}, & y_i \neq c. \end{cases}$$

Thus, we have shown that the objective of the dual can be rewritten in terms of variables $\{\beta_{c,i}\}$. After rewriting the constraints in terms of $\{\beta_{c,i}\}$, we have shown that the dual of the SVM (Equation (7)) can be equivalently written as in Equation (26). Note that the first constraint in (26) ensures consistency with the definition of $\boldsymbol{\beta}_c$ in Equation (31). The second constraint guarantees the non-negativity constraint of the original dual variables in (29), because we have

$$\beta_{c,i} z_{c,i} = \frac{\lambda_{c,i}}{k} \text{ for all } i \in [n], c \in [k] : c \neq y_i.$$

Consequently, we have

$$\beta_{c,i} z_{c,i} \geq 0 \iff \lambda_{c,i} \geq 0 \tag{33}$$

for all $c \in [k]$ and $i \in [n] : y_i \neq c$. In fact, the equivalence above also holds with the inequalities replaced by strict inequalities. Also note that the second constraint for $c = y_i$ yields $\frac{k-1}{k} \sum_{c' \neq y_i} \lambda_{c',i} \geq 0$, which is automatically satisfied when Equation (33) is satisfied. Thus, these constraints are redundant.

Step 2: Proof of Equation (13). Define

$$\hat{\boldsymbol{\beta}}_c := (\mathbf{X}^\top \mathbf{X})^+ \mathbf{z}_c, \ \forall c \in [k].$$

This specifies an *unconstrained* maximizer in (26). We will show that this unconstrained maximizer $\hat{\boldsymbol{\beta}}_c, c \in [k]$ is feasible in the constrained program in (26). Thus, it is in fact an optimal solution in (26).

To prove this, we will first prove that $\hat{\boldsymbol{\beta}}_c, c \in [k]$ satisfies the n equality constraints in (26). For convenience, let $\mathbf{g}_i \in \mathbb{R}^n, i \in [n]$ denote the i-th row of $(\mathbf{X}^\top \mathbf{X})^+$. Then, for the i-th element $\hat{\beta}_{c,i}$ of $\hat{\boldsymbol{\beta}}_c$, it holds that $\hat{\beta}_{c,i} = \mathbf{g}_i^\top \mathbf{z}_c$. Thus, for all $i \in [n]$, we have

$$\hat{\beta}_{y_i,i} + \sum_{c \neq y_i} \hat{\beta}_{c,i} = \mathbf{g}_i^\top \Big(\mathbf{z}_{y_i} + \sum_{c \neq y_i} \mathbf{z}_c \Big) = \mathbf{g}_i^\top \Big(\sum_{c \in [k]} \mathbf{z}_c \Big) = 0,$$

where in the last equality we used the definition of $\mathbf{z}_c$ in (11) and the fact that $\sum_{c \in [k]} \mathbf{v}_c = \mathbf{1}_n$, since each column of the label matrix $\mathbf{Y}$ has exactly one non-zero element equal to 1. Second, since Equation (12) holds, $\hat{\boldsymbol{\beta}}_c, c \in [k]$ further satisfies the n *strict* inequality constraints in (26).

We have shown that the unconstrained maximizer is feasible in the constrained program (26). Thus, we can conclude that it is also a global solution to the latter. By Equation (33), we note that the corresponding original dual variables $\{\hat{\lambda}_{c,i}\} = \{k\, \hat{\beta}_{c,i} z_{c,i}\}$ are all strictly positive. Now recall that under strong duality, any pair of primal-dual optimal solutions satisfies the KKT conditions. Hence the primal-dual pair $\big(\{\hat{\mathbf{w}}_c\}, \{\hat{\lambda}_{c,i}\}\big)$ satisfies the complementary slackness condition of Equation (30). This together with the positivity of $\{\hat{\lambda}_{c,i}\}$ complete the proof of the first part of the theorem, i.e. the proof of Equation (13).

Step 3: Proof of Equation (14). To prove Equation (14), consider the following OvA-type classifier: for all $c \in [k]$,

$$\min_{\mathbf{w}_c} \ \frac{1}{2}\|\mathbf{w}_c\|_2^2 \qquad \text{sub. to} \quad \mathbf{x}_i^\top \mathbf{w}_c \begin{cases} \geq \frac{k-1}{k}, & y_i = c, \\ \leq -\frac{1}{k}, & y_i \neq c, \end{cases} \quad \forall i \in [n]. \tag{34}$$

To see the connection with Equation (14), note the condition for the constraints in (34) to be active is exactly Equation (14). Thus, it suffices to prove that the constraints of (34) are active under the theorem's assumptions. We work again with the dual of (34):

$$\max_{\boldsymbol{\nu}_c \in \mathbb{R}^k} \ -\frac{1}{2}\|\mathbf{X}\boldsymbol{\nu}_c\|_2^2 + \mathbf{z}_c^\top \boldsymbol{\nu}_c \qquad \text{sub. to} \quad \mathbf{z}_c \odot \boldsymbol{\nu}_c \geq \mathbf{0}. \tag{35}$$

Again by complementary slackness, the desired Equation (14) holds provided that all dual constraints in (35) are strict at the optimal.

We now observe two critical similarities between (35) and (26): (i) the two dual problems have the same objectives (indeed the objective in (26) is separable over $c \in [k]$); (ii) they share the constraint $\mathbf{z}_c \odot \boldsymbol{\nu}_c \geq \mathbf{0}$ / $\mathbf{z}_c \odot \boldsymbol{\beta}_c \geq \mathbf{0}$. From this observation, we can use the same argument as for (26) to show that when Equation (12) holds, $\hat{\boldsymbol{\beta}}_c$ is optimal in (35).

Now, let $\mathrm{OPT}_{(28)}$ and $\mathrm{OPT}^{\mathrm{c}}_{(34)}$ be the optimal costs of the multiclass SVM in (28) and of the simplex-type OvA-SVM in (34) parameterized by $c \in [k]$. Also, denote $\mathrm{OPT}_{(26)}$ and $\mathrm{OPT}^{\mathrm{c}}_{(35)}, \mathrm{c} \in [\mathrm{k}]$ the optimal costs of their respective duals in (26) and (35), respectively. We proved above that

$$\mathrm{OPT}_{(26)} = \sum_{\mathrm{c} \in [\mathrm{k}]} \mathrm{OPT}^{\mathrm{c}}_{(35)}. \tag{36}$$

Further let $\mathbf{W}_{\text{OvA}} = [\mathbf{w}_{\text{OvA},1}, \ldots, \mathbf{w}_{\text{OvA},k}]$ be the optimal solution in the simplex-type OvA-SVM in (35). We have proved that under Equation (12) $\mathbf{w}_{\text{OvA},c}$ satisfies the constraints in (34) with equality, that is $\mathbf{X}^\top \mathbf{w}_{\text{OvA},c} = \mathbf{z}_c,\ \forall c \in [k]$. Thus, it suffices to prove that $\mathbf{W}_{\text{OvA}} = \mathbf{W}_{\text{SVM}}$. By strong duality (which holds trivially for (34) by Slater's conditions), we get

$$\begin{aligned}
\text{OPT}^{\text{c}}_{(34)} = \text{OPT}^{\text{c}}_{(35)},\ \text{c} \in [\text{k}] \implies & \sum_{c\in[k]} \text{OPT}^{\text{c}}_{(34)} = \sum_{\text{c}\in[\text{k}]} \text{OPT}^{\text{c}}_{(35)} \\
\overset{(36)}{\implies} & \sum_{c\in[k]} \text{OPT}^{\text{c}}_{(34)} = \text{OPT}_{(26)} \\
\overset{(34)}{\implies} & \sum_{c\in[k]} \frac{1}{2}\|\mathbf{w}_{\text{OvA},c}\|_2^2 = \text{OPT}_{(26)}.
\end{aligned} \tag{37}$$

Again, by strong duality we get $\text{OPT}_{(26)} = \text{OPT}_{(28)}$. Thus, we have

$$\sum_{c\in[k]} \frac{1}{2}\|\mathbf{w}_{\text{OvA},c}\|_2^2 = \text{OPT}_{(28)}.$$

Note also that $\mathbf{W}_{\text{OvA}}$ is feasible in (28) since

$$\mathbf{X}^\top \mathbf{w}_{\text{OvA},c} = \mathbf{z}_c,\ \forall c \in [k] \implies (\mathbf{w}_{\text{OvA},y_i} - \mathbf{w}_{\text{OvA},c})^\top \mathbf{x}_i = 1,\ \forall c \neq y_i, c \in [k],\ \text{and } \forall i \in [n].$$

Therefore, $\mathbf{W}_{\text{OvA}}$ is optimal in (28). Finally, note that the optimization objective in (28) is strongly convex. Thus, it has a unique minimum and therefore $\mathbf{W}_{\text{SVM}} = \mathbf{W}_{\text{OvA}}$ as desired.

6.2 Proof of Theorem 2

In this section, we provide the proof of Theorem 2. First, we remind the reader of the prescribed approach outlined in Section 3.2.1 and introduce some necessary notation. Second, we present the key Lemma 2, which forms the backbone of our proof. The proof of the lemma is rather technical and is deferred to Appendix A.1 along with a series of auxiliary lemmas. Finally, we end this section by showing how to prove Theorem 2 using Lemma 2.

Argument sketch and notation. We begin by presenting high-level ideas and defining notation that is specific to this proof. For $c \in [k]$, we define

$$\mathbf{A}_c := (\mathbf{Q} + \sum_{j=1}^{c} \boldsymbol{\mu}_j \mathbf{v}_j^T)^T (\mathbf{Q} + \sum_{j=1}^{c} \boldsymbol{\mu}_j \mathbf{v}_j^T).$$

Recall that in the above, $\boldsymbol{\mu}_j$ denotes the j^{th} class mean of dimension p, and $\mathbf{v}_j$ denotes the n-dimensional indicator that each training example is labeled as class j. Further, recall from Equation (1) that the feature matrix can be expressed as $\mathbf{X} = \mathbf{M}\mathbf{Y} + \mathbf{Q}$, where $\mathbf{Q} \in \mathbb{R}^{p\times n}$ is a standard Gaussian matrix. Thus, we have

$$\mathbf{X}^T\mathbf{X} = \mathbf{A}_k \qquad \text{and} \qquad \mathbf{Q}^T\mathbf{Q} = \mathbf{A}_0.$$

As discussed in Section 3.2.1, our goal is to show that the inverse Gram matrix $\mathbf{A}_k^{-1}$ is "close" to a positive definite diagonal matrix. Indeed, in our new notation, the desired inequality in Equation (12) becomes

$$z_{ci}\mathbf{e}_i^T \mathbf{A}_k^{-1} \mathbf{z}_c > 0, \quad \text{for all} \quad c \in [k] \quad \text{and } i \in [n]. \tag{38}$$

The major challenge in showing inequality (38) is that $\mathbf{A}_k = (\mathbf{Q} + \sum_{j=1}^{k} \boldsymbol{\mu}_j \mathbf{v}_j^T)^T (\mathbf{Q} + \sum_{j=1}^{k} \boldsymbol{\mu}_j \mathbf{v}_j^T)$ involves multiple mean components through the sum $\sum_{j=1}^{c} \boldsymbol{\mu}_j \mathbf{v}_j^T$. This makes it challenging to bound quadratic forms involving the Gram matrix $\mathbf{A}_k^{-1}$ directly. Instead, our idea is to work recursively

starting from bounding quadratic forms involving $\mathbf{A}_0^{-1}$. Specifically, we denote $\mathbf{P}_1 = \mathbf{Q} + \boldsymbol{\mu}_1\mathbf{v}_1^T$ and derive the following recursion on the $\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_k$ matrices:

$$\mathbf{A}_1 = \mathbf{P}_1^T\mathbf{P}_1 = \mathbf{A}_0 + \begin{bmatrix}\|\boldsymbol{\mu}_1\|_2\mathbf{v}_1 & \mathbf{Q}^T\boldsymbol{\mu}_1 & \mathbf{v}_1\end{bmatrix}\begin{bmatrix}\|\boldsymbol{\mu}_1\|_2\mathbf{v}_1^T \\ \mathbf{v}_1^T \\ \boldsymbol{\mu}_1^T\mathbf{Q}\end{bmatrix},$$

$$\mathbf{A}_2 = (\mathbf{P}_1 + \boldsymbol{\mu}_2\mathbf{v}_2^T)^T(\mathbf{P}_1 + \boldsymbol{\mu}_2\mathbf{v}_2^T) = \mathbf{A}_1 + \begin{bmatrix}\|\boldsymbol{\mu}_2\|_2\mathbf{v}_2 & \mathbf{P}_1^T\boldsymbol{\mu}_2 & \mathbf{v}_2\end{bmatrix}\begin{bmatrix}\|\boldsymbol{\mu}_2\|_2\mathbf{v}_2^T \\ \mathbf{v}_2^T \\ \boldsymbol{\mu}_2^T\mathbf{P}_1\end{bmatrix}, \tag{39}$$

and so on, until $\mathbf{A}_k$ (see Appendix F.1 for the complete expressions for the recursion). Using this trick, we can exploit bounds on quadratic forms involving $\mathbf{A}_0^{-1}$ to obtain bounds for quadratic forms involving $\mathbf{A}_1^{-1}$, and so on until $\mathbf{A}_k^{-1}$. Note that because of the nearly equal-energy Assumption 1, the order of adding mean vectors in $\mathbf{A}$ will not change the results. In other words, including $\boldsymbol{\mu}_1$ first, then $\boldsymbol{\mu}_2$, in (39) will produce the same result as including $\boldsymbol{\mu}_2$ first, then $\boldsymbol{\mu}_1$ in the same equation.

There are two key ideas behind this approach. First, we will show how to use a leave-one-out argument and the Matrix Inversion Lemma to express (recursively) the quadratic form $\mathbf{e}_i^T\mathbf{A}_k^{-1}\mathbf{z}_c$ in (38) in terms of simpler quadratic forms, which are more accessible to bound directly. For later reference, we define these auxiliary forms here. Let $\mathbf{d}_c := \mathbf{Q}^T\boldsymbol{\mu}_c$, for $c \in [k]$ and define the following quadratic forms involving $\mathbf{A}_c^{-1}$ for $c, j, m \in [k]$ and $i \in [n]$:

$$\begin{aligned} s_{mj}^{(c)} &:= \mathbf{v}_m^T\mathbf{A}_c^{-1}\mathbf{v}_j, \\ t_{mj}^{(c)} &:= \mathbf{d}_m^T\mathbf{A}_c^{-1}\mathbf{d}_j, \\ h_{mj}^{(c)} &:= \mathbf{v}_m^T\mathbf{A}_c^{-1}\mathbf{d}_j, \\ g_{ji}^{(c)} &:= \mathbf{v}_j^T\mathbf{A}_c^{-1}\mathbf{e}_i, \\ f_{ji}^{(c)} &:= \mathbf{d}_j^T\mathbf{A}_c^{-1}\mathbf{e}_i. \end{aligned} \tag{40}$$

For convenience, we refer to terms above as *quadratic forms of order c* or *the c-th order quadratic forms*, where c indicates the corresponding superscript. A complementary useful observation facilitating our approach is the observation that the class label indicators are orthogonal by definition, i.e. $\mathbf{v}_i^T\mathbf{v}_j = 0$, for $i, j \in [k]$. (This is a consequence of the fact that any training data point has a unique label and we are using here one-hot encoding.) Thus, the newly added mean component $\boldsymbol{\mu}_{c+1}\mathbf{v}_{c+1}^T$ is orthogonal to the already existing mean components included in the matrix $\mathbf{A}_c$ (see Equation (39)). Consequently, we will see that adding new mean components will only slightly change the magnitude of these these quadratic forms as c ranges from 0 to k.

Identifying and bounding quadratic forms of high orders. Recall the desired inequality (38). We can equivalently write the definition of $\mathbf{z}_c$ in Equation (11) as

$$\mathbf{z}_c = \frac{k-1}{k}\mathbf{v}_c + \sum_{j\neq c}\left(-\frac{1}{k}\right)\mathbf{v}_j = \tilde{z}_{c(c)}\mathbf{v}_c + \sum_{j\neq c}\tilde{z}_{j(c)}\mathbf{v}_j, \tag{41}$$

where we denote

$$\tilde{z}_{j(c)} = \begin{cases} -\frac{1}{k}, & \text{if } \ j \neq c \\ \frac{k-1}{k}, & \text{if } \ j = c \end{cases}.$$

Note that by this definition, we have $\tilde{z}_{y_i(c)} := z_{ci}$. This gives us

$$\begin{aligned} z_{ci}\mathbf{e}_i^T\mathbf{A}_k^{-1}\mathbf{z}_c &= z_{ci}^2\mathbf{e}_i^T\mathbf{A}_k^{-1}\mathbf{v}_{y_i} + \sum_{j\neq y_i} z_{ci}\tilde{z}_{j(c)}\mathbf{e}_i^T\mathbf{A}_k^{-1}\mathbf{v}_j, \\ &= z_{ci}^2 g_{y_i i}^{(k)} + \sum_{j\neq y_i} z_{ci}\tilde{z}_{j(c)}g_{ji}^{(k)}. \end{aligned} \tag{42}$$

Note that this expression (Equation (42)) involves the k-th order quadratic forms $g_{ji}^{(k)} = \mathbf{e}_i^T \mathbf{A}_k^{-1} \mathbf{v}_j$. For each such form, we use the matrix inversion lemma to leave the j-th mean component in $\mathbf{A}_k$ out and express it in terms of the *leave-one-out* versions of quadratic forms that we defined in (40), as below (see Appendix F.1 for a detailed derivation):

$$g_{ji}^{(k)} = \mathbf{e}_i^T \mathbf{A}_k^{-1} \mathbf{v}_j = \frac{(1 + h_{jj}^{(-j)}) g_{ji}^{(-j)} - s_{jj}^{(-j)} f_{ji}^{(-j)}}{s_{jj}^{(-j)} (\|\boldsymbol{\mu}_j\|_2^2 - t_{jj}^{(-j)}) + (1 + h_{jj}^{(-j)})^2}. \tag{43}$$

Specifically, above we defined $s_{jj}^{(-j)} := \mathbf{v}_j^T \mathbf{A}_{-j}^{-1} \mathbf{v}_j$, where $\mathbf{A}_{-j}$ denotes the version of the Gram matrix $\mathbf{A}_k$ with the j-th mean component left out. The quadratic forms $h_{jj}^{(-j)}$, $f_{ji}^{(-j)}$, $g_{ji}^{(-j)}$ and $t_{jj}^{(-j)}$ are defined similarly in view of Equation (40).

Specifically, to see how these "leave-one-out" quadratic forms relate directly to the forms in Equation (40), note that it suffices in (43) to consider the case where $j = k$. Indeed, observe that when $j \neq k$ we can simply change the order of adding mean components, described in Equation (39), so that the j-th mean component is added last. On the other hand, when $j = k$ the leave-one-out quadratic terms in (43) involve the Gram matrix $\mathbf{A}_{k-1}$. Thus, they are equal to the quadratic forms of order $k-1$, given by $s_{kk}^{(k-1)}, t_{kk}^{(k-1)}$, $h_{kk}^{(k-1)}$, $g_{ki}^{(k-1)}$ and $f_{ki}^{(k-1)}$.

The following technical lemma bounds all of these quantities and its use is essential in the proof of Theorem 2. Its proof, which is deferred to Appendix A, relies on the recursive argument outlined above: We start from the quadratic forms of order 0 building up all the way to the quadratic forms of order $k-1$.

Lemma 2 (Quadratic forms of high orders)**.** *Let Assumption 1 hold and further assume that $p > Ck^3 n \log(kn) + n - 1$ for large enough constant $C > 1$ and large n. There exist constants c_i's and C_i's > 1 such that the following bounds hold for every $i \in [n]$ and $j \in [k]$ with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$,*

$$\begin{aligned}
\frac{C_1 - 1}{C_1} \cdot \frac{n}{kp} \leq & s_{jj}^{(-j)} \leq \frac{C_1 + 1}{C_1} \cdot \frac{n}{kp}, \\
t_{jj}^{(-j)} \leq & \frac{C_2 n \|\boldsymbol{\mu}\|_2^2}{p}, \\
-\tilde{\rho}_{n,k} \frac{C_3 n \|\boldsymbol{\mu}\|_2}{\sqrt{kp}} \leq & h_{jj}^{(-j)} \leq \tilde{\rho}_{n,k} \frac{C_3 n \|\boldsymbol{\mu}\|_2}{\sqrt{kp}}, \\
|f_{ji}^{(-j)}| \leq & \frac{C_4 \sqrt{n} \|\boldsymbol{\mu}\|_2}{p}, \\
g_{ji}^{(-j)} \geq & \left(1 - \frac{1}{C_5}\right) \frac{1}{p}, \quad \text{for } j = y_i, \\
|g_{ji}^{(-j)}| \leq & \frac{1}{C_6 k^2 p}, \quad \text{for } j \neq y_i,
\end{aligned}$$

where $\tilde{\rho}_{n,k} = \min\{1, \sqrt{\log(2n)/k}\}$. Observe that the bounds stated in the lemma hold for any $j \in [k]$ and the bounds themselves are independent of j.

Completing the proof of Theorem 2. We now show how to use Lemma 2 to complete the proof of the theorem. Following the second condition in the statement of Theorem 2, we define

$$\epsilon_n := \frac{k^{1.5} n \sqrt{n} \|\boldsymbol{\mu}\|_2}{p} \leq \tau, \tag{44}$$

where τ is a sufficiently small positive constant, the value of which will be specified later in the proof. First, we will show that the denominator of Equation (43) is strictly positive on the event where Lemma 2 holds. We define

$$\det\nolimits_{-j} := s_{jj}^{(-j)} (\|\boldsymbol{\mu}_j\|_2^2 - t_{jj}^{(-j)}) + (1 + h_{jj}^{(-j)})^2.$$

By Lemma 2, the quadratic forms $s_{jj}^{(-j)}$ are of the same order $\Theta\left(\frac{n}{kp}\right)$ for every $j \in [k]$. Similarly, we have $t_{jj}^{(-j)} = \mathcal{O}\left(\frac{n}{p}\|\boldsymbol{\mu}\|_2^2\right)$ and $|h_{jj}^{(-j)}| = \tilde{\rho}_{n,k}\mathcal{O}\left(\frac{\epsilon_n}{k^2\sqrt{n}}\right)$ for $j \in [k]$. Thus, we have

$$\frac{n\|\boldsymbol{\mu}\|_2^2}{C_1 kp}\left(1 - \frac{C_2 n}{p}\right) + \left(1 - \frac{C_3\epsilon_n}{k^2\sqrt{n}}\right)^2 \leq \det{}_{-j} \leq \frac{C_1 n\|\boldsymbol{\mu}\|_2^2}{kp} + \left(1 + \frac{C_3\epsilon_n}{k^2\sqrt{n}}\right)^2, \tag{45}$$

with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$, for every $j \in [k]$. Here, we use the fact that $t_{jj}^{-j} \geq 0$ by the positive semidefinite property of the leave-one-out Gram matrix $\mathbf{A}_{-j}^{-1}$. Next, we choose τ in Equation (44) to be sufficiently small so that $C_3\tau \leq 1/2$. Provided that p is sufficiently large compared to n, there then exist constants $C_1', C_2' > 0$ such that we have

$$C_1' \leq \frac{\det_{-m}}{\det_{-j}} \leq C_2', \quad \text{for all } j, m \in [k],$$

with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$. Now, assume without loss of generality that $y_i = k$. Equation (45) shows that there exists constant $c > 0$ such that $\det_{-j} > c$ for all $j \in [k]$ with high probability provided that p/n is large enough (guaranteed by the first condition of the theorem). Hence, to make the right-hand-side of Equation (42) positive, it suffices to show that the numerator will be positive. Accordingly, we will show that

$$z_{ci}^2\big((1 + h_{kk}^{(-k)})g_{ki}^{(-k)} - s_{kk}^{(-k)} f_{ki}^{(-k)}\big) + C z_{ci} \sum_{j \neq k} \tilde{z}_j \big((1 + h_{jj}^{(-j)})g_{ji}^{(-j)} - s_{jj}^{(-j)} f_{ji}^{(-j)}\big) > 0, \tag{46}$$

for some $C > 1$.

We can show by simple algebra that it suffices to consider the worst case of $z_{ci} = -1/k$. To see why this is true, we consider the simpler term $z_{ci}^2 g_{y_i i}^{(-y_i)} - |\sum_{j \neq y_i} z_{ci}\tilde{z}_{j(c)} g_{ji}^{(-j)}|$. Clearly, Equation (46) is positive only if the above quantity is also positive. Lemma 2 shows that when $z_{ci} = -1/k$, then $z_{ci}^2 g_{y_i i}^{(-y_i)} \geq \left(1 - \frac{1}{C_1}\right)\frac{1}{k^2 p}$ and $|z_{ci}\tilde{z}_{j(c)} g_{ji}^{(-j)}| \leq \frac{1}{C_2 k^3 p}$, for $j \neq y_i$. Hence

$$z_{ci}^2 g_{y_i i}^{(-y_i)} - |\sum_{j \neq y_i} z_{ci}\tilde{z}_{j(c)} g_{ji}^{(-j)}| \geq \left(1 - \frac{1}{C_3}\right)\frac{1}{k^2 p}.$$

Here, $z_{ci} = -1/k$ minimizes the lower bound $z_{ci}^2 g_{y_i i}^{(-y_i)} - |\sum_{j \neq y_i} z_{ci}\tilde{z}_{j(c)} g_{ji}^{(-j)}|$. To see this, we first drop the positive common factor $|z_{ci}|$ in the equation above and get $|z_{ci}| g_{y_i i}^{(-y_i)} - |\sum_{j \neq y_i} \tilde{z}_{j(c)} g_{ji}^{(-j)}|$. If we had $z_{ci} = -1/k$, then $|\tilde{z}_{j(c)}|$ is either $(k-1)/k$ or $1/k$. In contrast, if we consider $z_{ci} = (k-1)/k$, then we have $|\tilde{z}_{j(c)}| = 1/k$ for all $j \neq y_i$ and so the term $|z_{ci}| g_{y_i i}^{(-y_i)} - |\sum_{j \neq y_i} \tilde{z}_{j(c)} g_{ji}^{(-j)}|$ is strictly larger.

Using this worst case, i.e. $z_{ci} = -1/k$, and the trivial inequality $|\tilde{z}_{j(c)}| < 1$ for $j \neq y_i$ together with the bounds for the terms $s_{jj}^{(-j)}, t_{jj}^{(-j)}, h_{jj}^{(-j)}$ and $f_{ji}^{(-j)}$ derived in Lemma 2 gives us

$$\begin{aligned}
(46) &\geq \frac{1}{k^2}\left(\left(1 - \frac{C_1\epsilon_n}{k^2\sqrt{n}}\right)\left(1 - \frac{1}{C_2}\right)\frac{1}{p} - \frac{C_3\epsilon_n}{k^{1.5}n}\cdot\frac{n}{kp}\right) - k\cdot\frac{1}{C_4 k}\left(\left(1 + \frac{C_5\epsilon_n}{k^2\sqrt{n}}\right)\frac{1}{k^2 p} - \frac{C_6\epsilon_n}{k^{1.5}n}\frac{n}{kp}\right) \\
&\geq \frac{1}{k^2}\left(1 - \frac{1}{C_9} - \frac{C_{10}\epsilon_n}{k^2\sqrt{n}} - \frac{C_{11}\epsilon_n}{k^2} - C_{12}\epsilon_n\right)\frac{1}{p} \\
&\geq \frac{1}{k^2 p}\left(1 - \frac{1}{C_9} - C_{10}\tau\right),
\end{aligned} \tag{47}$$

with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$ for some constants C_i's > 1. Above, we recalled the definition of ϵ_n and used from Lemma 2 that $h_{jj}^{(-j)} \leq \tilde{\rho}_{n,k}\frac{C_{11}\epsilon_n}{k^2\sqrt{n}}$ and $|f_{ji}^{(-j)}| \leq \frac{C_{12}\epsilon_n}{k^{1.5}n}$ with high probability. To complete the proof, we choose τ to be a small enough constant to guarantee $C_{10}\tau < 1 - 1/C_9$, and substitute this in Equation (47) to get the desired condition of Equation (46).

6.3 Proof of Theorem 5

Challenges and notation. We begin by highlighting the two main non-trivialities introduced in the analysis of the multiclass setting. We compare them to the binary-error analysis in [MNS+21] and we sketch our approach to each one of them:

- *The multitude of signal vectors:* The generative model for the MLM involves k distinct (high-dimensional) signal vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$, and the classification error is a complicated functional of all k *recovered* signal vectors (denoted by $\widehat{\mathbf{w}}_1, \ldots, \widehat{\mathbf{w}}_k$ respectively). This functional has to be dealt with carefully compared to the binary case, where there is only one signal vector. In particular, direct plug-ins of the *survival signal* and *contamination factor* for each recovered signal vector (here, we follow the terminology in [MVSS20]) do not provide sufficiently sharp expressions of the multiclass classification error to predict separation between classification-consistency and regression-consistency. We circumvent this issue by directly analyzing survival and contamination factors of the pairwise *difference signal* between two classes, and showing in Lemmas 4 and 5 that they scale very similarly to the single-signal case. We note that while the survival and contamination factors of this pairwise difference signal scale identically to the single-signal case, the proofs do not follow as a corollary of the corresponding lemmas in [MNS+21]; in particular, the difference of label vectors turns out to depend not only on a single "difference" feature but all the top k features. This requires a much more complex leave-k-out analysis, as opposed to the simpler leave-one-out analysis carried out in [BLLT20, MNS+21].
- *Covariate-dependent label noise in the MLM:* The error analysis provided in [MNS+21] critically leverages that the cross-correlation between the logit and the binary label of a training example is lower bounded by a universal positive constant. This is relatively straightforward to show when the relationship between the logit and the label is one of *constant-label-noise* where the event of label error is independent of the covariate. On the other hand, the MLM involves label errors that are highly depend on the covariate, and these cross-correlation terms need to be handled much more carefully. We provide an elegant argument based on Stein's lemma to handle the more complex MLM-induced label noise.

Before proceeding we set up some notation for important quantities in the analysis. Note that Assumption 5 directly implies that $\mu_{c,j_c} = 1$ for all $c \in [k]$. For any two classes $c_1 \neq c_2$, we define the true *difference signal vector* as

$$\boldsymbol{\Delta}_{c_1,c_2} := \boldsymbol{\mu}_{c_1} - \boldsymbol{\mu}_{c_2} = \mu_{c_1,j_{c_1}} \boldsymbol{e}_{j_{c_1}} - \mu_{c_2,j_{c_2}} \boldsymbol{e}_{j_{c_2}},$$

where the last step follows from Assumption 5. Correspondingly, the recovered difference signal vector is defined as $\widehat{\boldsymbol{\Delta}}_{c_1,c_2} := \widehat{\mathbf{w}}_{c_1} - \widehat{\mathbf{w}}_{c_2}$.

Identifying the survival and contamination terms. We state and prove our main lemma that characterizes the classification error in MLM as a function of effective survival and contamination terms.

Lemma 3. *The excess classification risk is bounded by*

$$\mathbb{P}_e - \mathbb{P}_{e,\mathsf{Bayes}} \leq \sum_{c_1<c_2} \left(\frac{1}{2} - \frac{1}{\pi} \mathsf{tan}^{-1} \left(\frac{\mathsf{SU}(\widehat{\boldsymbol{\Delta}}_{c_1,c_2}, \boldsymbol{\Delta}_{c_1,c_2})}{\mathsf{CN}(\widehat{\boldsymbol{\Delta}}_{c_1,c_2}, \boldsymbol{\Delta}_{c_1,c_2})} \right) \right), \tag{48}$$

where we define for any two classes $c_1 \neq c_2 \in [k]$:

$$\mathsf{SU}(\widehat{\boldsymbol{\Delta}}_{c_1,c_2}, \boldsymbol{\Delta}_{c_1,c_2}) := \frac{\widehat{\boldsymbol{\Delta}}^\top_{c_1,c_2} \boldsymbol{\Sigma} \boldsymbol{\Delta}_{c_1,c_2}}{\|\boldsymbol{\Sigma}^{1/2} \boldsymbol{\Delta}_{c_1,c_2}\|_2} \text{ and}$$

$$\mathsf{CN}(\widehat{\boldsymbol{\Delta}}_{c_1,c_2}, \boldsymbol{\Delta}_{c_1,c_2}) := \sqrt{\left(\widehat{\boldsymbol{\Delta}}_{c_1,c_2} - \frac{\widehat{\boldsymbol{\Delta}}^\top_{c_1,c_2} \boldsymbol{\Sigma} \boldsymbol{\Delta}_{c_1,c_2}}{\|\boldsymbol{\Sigma}^{1/2} \boldsymbol{\Delta}_{c_1,c_2}\|_2^2} \boldsymbol{\Delta}_{c_1,c_2} \right)^\top \boldsymbol{\Sigma} \left(\widehat{\boldsymbol{\Delta}}_{c_1,c_2} - \frac{\widehat{\boldsymbol{\Delta}}^\top_{c_1,c_2} \boldsymbol{\Sigma} \boldsymbol{\Delta}_{c_1,c_2}}{\|\boldsymbol{\Sigma}^{1/2} \boldsymbol{\Delta}_{c_1,c_2}\|_2^2} \boldsymbol{\Delta}_{c_1,c_2} \right)}.$$

Proof. We consider a fixed $\mathbf{x}$, and (following the notation in [TOS20]) the k-dimensional vectors

$$\mathbf{g} := \begin{bmatrix} \mathbf{x}^\top \widehat{\mathbf{w}}_1 & \mathbf{x}^\top \widehat{\mathbf{w}}_2 & \dots & \mathbf{x}^\top \widehat{\mathbf{w}}_k \end{bmatrix}$$
$$\mathbf{h} := \begin{bmatrix} \mathbf{x}^\top \boldsymbol{\mu}_1 & \mathbf{x}^\top \boldsymbol{\mu}_2 & \dots & \mathbf{x}^\top \boldsymbol{\mu}_k \end{bmatrix}$$

Further, we define the *multinomial logit* variable $Y(\mathbf{h})$ such that

$$\mathbb{P}\left[Y(\mathbf{h}) = j\right] = \frac{\exp\{h_j\}}{\sum_{m=1}^{k} \exp\{h_m\}}.$$

Recall that $\mathbb{P}_e = \mathbb{P}\left(\arg\max(\mathbf{g}) \neq Y(\mathbf{h})\right)$, where the probability is taken both over the fresh test sample $\mathbf{x}$ and the randomness in the multinomial logit variable. We note that for there to be a classification error conditioned on $\mathbf{x}$, *at least* one of the following two events needs to hold: a) $\arg\max(\mathbf{g}) \neq \arg\max(\mathbf{h})$, or b) $Y(\mathbf{h}) \neq \arg\max(\mathbf{h})$. To see this, note that if neither a) nor b) held, we would have $\arg\max(\mathbf{g}) = Y(\mathbf{h})$ and we would not have a classification error conditional on the covariate being $\mathbf{x}$. Thus, applying a union bound gives us

$$\begin{aligned} \mathbb{P}_e &\leq \mathbb{P}_{e,0} + \mathbb{P}_{e,\mathsf{Bayes}} \text{ where} \\ \mathbb{P}_{e,0} &:= \mathbb{P}\left(\arg\max(\mathbf{g}) \neq \arg\max(\mathbf{h})\right) \text{ and} \\ \mathbb{P}_{e,\mathsf{Bayes}} &:= \mathbb{P}\left(\arg\max(\mathbf{h}) \neq Y(\mathbf{h})\right). \end{aligned}$$

Thus, it suffices to provide an upper bound on $\mathbb{P}_{e,0}$ as defined. We note that for there to be an error of the form $\arg\max(\mathbf{g}) \neq \arg\max(\mathbf{h})$, there *needs* to exist indices $c_1, c_2 \in [k]$ (whose choice can depend on $\mathbf{x}$) such that $\mathbf{x}^\top \boldsymbol{\mu}_{c_1} \geq \mathbf{x}^\top \boldsymbol{\mu}_{c_2}$ but $\mathbf{x}^\top \widehat{\mathbf{w}}_{c_1} < \mathbf{x}^\top \widehat{\mathbf{w}}_{c_2}$. In other words, we have

$$\begin{aligned} \mathbb{P}_{e,0} &\leq \mathbb{P}\left(\mathbf{x}^\top \boldsymbol{\mu}_{c_1} \geq \mathbf{x}^\top \boldsymbol{\mu}_{c_2} \text{ and } \mathbf{x}^\top \widehat{\mathbf{w}}_{c_1} < \mathbf{x}^\top \widehat{\mathbf{w}}_{c_2} \text{ for some } c_1 \neq c_2\right) \\ &\leq \sum_{c_1 \neq c_2} \mathbb{P}\left(\mathbf{x}^\top \boldsymbol{\mu}_{c_1} \geq \mathbf{x}^\top \boldsymbol{\mu}_{c_2} \text{ and } \mathbf{x}^\top \widehat{\mathbf{w}}_{c_1} < \mathbf{x}^\top \widehat{\mathbf{w}}_{c_2}\right) \\ &= \sum_{c_1 < c_2} \mathbb{P}\left(\mathbf{x}^\top \boldsymbol{\Delta}_{c_1,c_2} \cdot \mathbf{x}^\top \widehat{\boldsymbol{\Delta}}_{c_1,c_2} < 0\right). \end{aligned}$$

Now, we consider whitened versions of the difference signal vectors: $\boldsymbol{E}_{c_1,c_2} := \boldsymbol{\Sigma}^{1/2} \boldsymbol{\Delta}_{c_1,c_2}$, $\widehat{\boldsymbol{E}}_{c_1,c_2} := \boldsymbol{\Sigma}^{1/2} \widehat{\boldsymbol{\Delta}}_{c_1,c_2}$. We also define the generalized survival and contamination terms of the difference signal vector as

$$\mathsf{SU}(\widehat{\boldsymbol{\Delta}}_{c_1,c_2}, \boldsymbol{\Delta}_{c_1,c_2}) := \frac{\widehat{\boldsymbol{E}}_{c_1,c_2}^T \boldsymbol{E}_{c_1,c_2}}{\|\boldsymbol{E}_{c_1,c_2}\|_2}$$
$$\mathsf{CN}(\widehat{\boldsymbol{\Delta}}_{c_1,c_2}, \boldsymbol{\Delta}_{c_1,c_2}) := \sqrt{\|\widehat{\boldsymbol{E}}_{c_1,c_2}\|_2^2 - \frac{\left(\widehat{\boldsymbol{E}}_{c_1,c_2}^T \boldsymbol{E}_{c_1,c_2}\right)^2}{\|\boldsymbol{E}_{c_1,c_2}\|_2^2}}$$

Recall that $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Then, the rotational invariance property of the Gaussian distribution and Gaussian decomposition yields:

$$\begin{aligned} \mathbb{P}\left(\mathbf{x}^\top \boldsymbol{\Delta}_{c_1,c_2} \cdot \mathbf{x}^\top \widehat{\boldsymbol{\Delta}}_{c_1,c_2} < 0\right) &= \mathbb{P}_{\mathbf{G} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}\left(\mathbf{G}^\top \boldsymbol{E}_{c_1,c_2} \cdot \mathbf{G}^\top \widehat{\boldsymbol{E}}_{c_1,c_2} < 0\right) \\ &= \mathbb{P}_{\substack{G \sim \mathcal{N}(0,1) \\ H \sim \mathcal{N}(0,1)}}\left(\|\boldsymbol{E}_{c_1,c_2}\|_2 G \cdot \left(\mathsf{SU}(\widehat{\boldsymbol{\Delta}}_{c_1,c_2}, \boldsymbol{\Delta}_{c_1,c_2}) G + \mathsf{CN}(\widehat{\boldsymbol{\Delta}}_{c_1,c_2}, \boldsymbol{\Delta}_{c_1,c_2}) H\right) < 0\right) \\ &= \mathbb{P}_{\substack{G \sim \mathcal{N}(0,1) \\ H \sim \mathcal{N}(0,1)}}\left(\left(\mathsf{SU}(\widehat{\boldsymbol{\Delta}}_{c_1,c_2}, \boldsymbol{\Delta}_{c_1,c_2})\, G^2 + \mathsf{CN}(\widehat{\boldsymbol{\Delta}}_{c_1,c_2}, \boldsymbol{\Delta}_{c_1,c_2})\, H\, G\right) < 0\right) \\ &= \frac{1}{2} - \frac{1}{\pi} \mathsf{tan}^{-1}\left(\frac{\mathsf{SU}(\widehat{\boldsymbol{\Delta}}_{c_1,c_2}, \boldsymbol{\Delta}_{c_1,c_2})}{\mathsf{CN}(\widehat{\boldsymbol{\Delta}}_{c_1,c_2}, \boldsymbol{\Delta}_{c_1,c_2})}\right). \end{aligned} \tag{49}$$

For the last equality in Equation (49), we used the fact that the ratio H/G of two independent standard normals follows the standard Cauchy distribution. This completes the proof. □

Bounding the survival and contamination terms. Next, we provide characterizations of $\mathsf{SU}(\widehat{\boldsymbol{\Delta}}_{c_1,c_2}, \boldsymbol{\Delta}_{c_1,c_2})$ and $\mathsf{CN}(\widehat{\boldsymbol{\Delta}}_{c_1,c_2}, \boldsymbol{\Delta}_{c_1,c_2})$. We abbreviate these by SU_{c_1,c_2} and CN_{c_1,c_2} respectively for brevity. These characterizations address two new aspects of the MLM: the multiclass setting, and label noise generated by the logistic model. We start with the characterization of survival.

Lemma 4 (Survival terms)**.** *There exist positive universal constants L_1, L_2, U_1, U_2, C such that*

$$\mathsf{SU}^L(n) \le \mathsf{SU}_{c_1,c_2}(n) \le \mathsf{SU}^U(n), \quad \textit{where}$$

$$\mathsf{SU}^L(n) := \begin{cases} c_k(1 + L_1 n^{q-(1-r)})^{-1}, & 0 < q < 1-r \\ c_k L_2 n^{(1-r)-q}, & q > 1-r. \end{cases}$$

$$\mathsf{SU}^U(n) := \begin{cases} c_k(1 + U_1 n^{q-(1-r)})^{-1}, & 0 < q < 1-r \\ c_k U_2 n^{(1-r)-q}, & q > 1-r. \end{cases}$$

with probability at least $1 - Ck^3 e^{-C\sqrt{n}}$. Above, $c_k > 0$ is a fixed strictly positive constant that depends on k but not on n.

Lemma 4 constitutes a nontrivial extension of Lemma 11 of [MNS+21] to deal with intricacies in the new pairwise-difference signal vector and the covariate-dependent label noise induced by the MLM. Its proof is provided in Appendix D.1.

Next, we provide an upper-bound characterization of contamination.

Lemma 5 (Contamination terms)**.** *There exists a universal constant C_k that depends only on k such that*

$$\mathsf{CN}_{c_1,c_2}(n) \le C_k \sqrt{\log n} \cdot n^{-\frac{\min\{m-1, 2q+r-1, 2q+2r-3/2\}}{2}}, q > 1-r$$

with probability at least $1 - \frac{C_k}{n^c}$ for some constant $0 < c \le 1$.

Lemma 5 extends Lemma 13 of [MNS+21] for binary classification, and its proof is provided in Appendix D.2. As with the analysis of survival, the dependency of the label difference vector on the top k features requires an intricate leave-k-out analysis Accordingly, several technical lemmas established in the proof of Lemma 4 are also used in this proof.

Plugging Lemmas 4 and 5 into Lemma 3 directly gives us the desired statement of Theorem 5. □

7 Conclusion and future work

Our work provides, to the best of our knowledge, the first results characterizing a) equivalence of loss functions, and b) generalization of interpolating solutions in multiclass settings. We outline here some immediate as well as longer-term future directions. First, in Section 4.1.1, we discussed in detail the potential for extending our techniques to anisotropic scenarios for GMM data. However, the formal details of such extensions require further work that is beyond the scope of this paper. Another important area for future research is the extension of our results to situations where the number of classes (k) scales with the problem dimensions (n, p). This is particularly intriguing as past research (e.g. [AGL21]) has shown, albeit under differing assumptions and with distinct training algorithms, that there is a different generalization error behavior between small and large numbers of classes. Despite our research's focus on the condition where k is constant, our results provide a mathematical basis for such extensions. A key contribution of our work is the establishment of deterministic equivalence conditions between multiclass SVM and MNI, which not only remain valid but also serve as a basis for analyzing any probabilistic data model and any scaling regime of k. In fact, after the initial release of this paper, the authors of [SAS22, WS23] leveraged our equivalence result and expanded our generalization bounds for the case of MLM data to the case where k can grow with n and p, which requires new technical insights.

More generally, our fine-grained techniques are tailored to high-dimensional linear models with Gaussian features. Furthermore, we believe the results derived here can extend to kernel machines and other nonlinear settings; formally showing these extensions is of substantial interest. It is also

interesting to investigate corresponding lower bounds for our results — for example, studying the sharpness of our conditions for equivalence of SVM to MNI in Section 3.2, analogous to [ASH21] for the binary case. Also, we have limited attention to balanced datasets throughout, i.e. we assumed that each class contains equal number of training samples. We would like to investigate the effect of data imbalances on our results extending our analysis to CE modifications tailored to imbalanced data recently proposed in [CWG+19, MJR+20, KPOT21]. Finally, we have established a tight connection of our findings regarding the geometry of support vectors under overparameterization with the neural collapse phenomenon. Nevertheless, many questions remain open towards better explaining what leads the learnt feature representations of overparameterized to have the observed ETF structure. It is a fascinating research direction further exploring the geometry of learnt features and of support vectors in nonlinear settings.

Acknowledgments

We are grateful to Vignesh Subramanian, Rahul Arya and Anant Sahai for pointing out a subtle issue in the proofs of Lemmas 4 and 5, which has since been fixed. We are also grateful to the anonymous reviewers for their valuable feedback, which has contributed to enhancing the presentation of the initial submission. This work is partially supported by the NSF under Grant Number CCF-2009030, by an NSERC Discovery Grant and by a grant from KAUST. Part of this work was done when VM was visiting the Simons Institute for the Theory of Computing.

References

[AGL21] Felix Abramovich, Vadim Grinshtein, and Tomer Levy. Multiclass classification by sparse multinomial logistic regression. *IEEE Transactions on Information Theory*, 67(7):4637–4646, 2021.

[AKLZ20] Benjamin Aubin, Florent Krzakala, Yue Lu, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching Bayes error with convex optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12199–12210. Curran Associates, Inc., 2020.

[ASH21] Navid Ardeshir, Clayton Sanford, and Daniel J Hsu. Support vector machines and linear regression coincide with very high-dimensional features. *Advances in Neural Information Processing Systems*, 34, 2021.

[ASS01] Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, September 2001.

[BB99] Erin J Bredensteiner and Kristin P Bennett. Multicategory classification by support vector machines. In *Computational Optimization*, pages 53–79. Springer, 1999.

[BEH20] Anna Sergeevna Bosman, Andries Engelbrecht, and Mardé Helbig. Visualising basins of attraction for the cross-entropy and the squared error neural network loss functions. *Neurocomputing*, 400:113–136, 2020.

[Ber09] Dennis S Bernstein. *Matrix mathematics: theory, facts, and formulas*. Princeton university press, 2009.

[BG01] Arnaud Buhot and Mirta B Gordon. Robust learning and generalization with support vector machines. *Journal of Physics A: Mathematical and General*, 34(21):4377–4388, May 2001.

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[BHX20] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[BLLT20] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[BM94] Kristin P. Bennett and O.L. Mangasarian. Multicategory discrimination via linear programming. *Optimization Methods and Software*, 3(1-3):27–39, 1994.

[BM03] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, March 2003.

[CGB21] Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-Gaussian mixtures. *arXiv preprint arXiv:2104.13628*, 2021.

[CKMY16] Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Structured prediction theory based on factor graph complexity. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[CL21] Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.

[CLRS09] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009.

[CS02] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, March 2002.

[CWG+19] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1567–1578, 2019.

[DB95] Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2(1):263–286, January 1995.

[DCO20] Ahmet Demirkaya, Jiasi Chen, and Samet Oymak. Exploring the role of loss functions in multiclass classification. In *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–5, 2020.

[DET05] David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2005.

[DKT21] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, April 2021.

[DL20] Oussama Dhifallah and Yue M Lu. A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*, 2020.

[DOS99] Rainer Dietrich, Manfred Opper, and Haim Sompolinsky. Statistical mechanics of support vector networks. *Physical Review Letters*, 82:2975–2978, Apr 1999.

[DR17] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[EHN96] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[FÖ2] Johannes Fürnkranz. Round robin classification. *Journal of Machine Learning Research*, 2:721–747, March 2002.

[FHLS21a] Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43), 2021.

[FHLS21b] Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43), 2021.

[GCOZ17] Krzysztof Gajowniczek, Leszek J. Chmielewski, Arkadiusz Orłowski, and Tomasz Ząbkowski. Generalized entropy cost function in neural networks. In Alessandra Lintas, Stefano Rovetta, Paul F.M.J. Verschure, and Alessandro E.P. Villa, editors, *Artificial Neural Networks and Machine Learning – ICANN 2017*, pages 128–136, Cham, 2017. Springer International Publishing.

[GHNK21] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021.

[GHST05] Thore Graepel, Ralf Herbrich, and John Shawe-Taylor. Pac-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.

[GJS+20] Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, February 2020.

[GLL+11] Pascal Germain, Alexandre Lacoste, François Laviolette, Mario Marchand, and Sara Shanian. A pac-bayes sample-compression approach to kernel methods. In *ICML*, 2011.

[HB20] Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

[HJ12] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, USA, 2nd edition, 2012.

[HMRT19] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[HMX21] Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 91–99. PMLR, 13–15 Apr 2021.

[HPD21] XY Han, Vardan Papyan, and David L Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021.

[Hua17] Hanwen Huang. Asymptotic behavior of support vector machine for spiked population model. *Journal of Machine Learning Research*, 18(45):1–21, 2017.

[HYS16] Le Hou, Chen-Ping Yu, and Dimitris Samaras. Squared earth mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*, 2016.

[IMSV19] Arya Iranmehr, Hamed Masnadi-Shirazi, and Nuno Vasconcelos. Cost-sensitive support vector machines. *Neurocomputing*, 343:50–64, 2019.

[JT19] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1772–1798, Phoenix, USA, 25–28 Jun 2019. PMLR.

[KA21] Abla Kammoun and Mohamed-Slim Alouini. On the precise error analysis of support vector machines. *IEEE Open Journal of Signal Processing*, 2:99–118, 2021.

[KLS20] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020.

[KP02] V. Koltchinskii and D. Panchenko. Empirical Margin Distributions and Bounding the Generalization Error of Combined Classifiers. *The Annals of Statistics*, 30(1):1 – 50, 2002.

[KPOT21] Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34, 2021.

[KS18] Himanshu Kumar and P. S. Sastry. Robust loss functions for learning multi-class classifiers. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 687–692, 2018.

[KT21] Ganesh Ramachandra Kini and Christos Thrampoulidis. Phase transitions for one-vs-one and one-vs-all linear separability in multiclass gaussian mixtures. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4020–4024, 2021.

[KTW+20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[LDBK15] Yunwen Lei, Urun Dogan, Alexander Binder, and Marius Kloft. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[LDZK19] Yunwen Lei, Ürün Dogan, Ding-Xuan Zhou, and Marius Kloft. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5):2995–3021, 2019.

[LLW04] Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

[Lol20] Panagiotis Lolas. Regularization in high-dimensional regression and classification via random matrix theory. *arXiv preprint arXiv:2003.13723*, 2020.

[LR21] Tengyuan Liang and Benjamin Recht. Interpolating classifiers make few mistakes. *arXiv preprint arXiv:2101.11815*, 2021.

[LS20] Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and min-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*, 2020.

[LS22] Jianfeng Lu and Stefan Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 2022.

[Mau16] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In Ronald Ortner, Hans Ulrich Simon, and Sandra Zilles, editors, *Algorithmic Learning Theory*, pages 3–17, Cham, 2016. Springer International Publishing.

[MJR+20] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2020.

[MLC19] Xiaoyi Mai, Zhenyu Liao, and Romain Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3361, 2019.

[MNS+21] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.

[MO05] Dörthe Malzahn and Manfred Opper. A statistical physics approach for the analysis of machine learning algorithms on real data. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):P11001–P11001, nov 2005.

[MPP20] Dustin G Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *arXiv preprint arXiv:2011.11619*, 2020.

[MR16] Yu Maximov and Daria Reshetova. Tight risk bounds for multi-class margin classifiers. *Pattern Recognition and Image Analysis*, 26:673–680, 2016.

[MRSY19] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[MVSS20] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 2020.

[PGS13] Bernardo Ávila Pires, Mohammad Ghavamzadeh, and Csaba Szepesvári. Cost-sensitive multiclass classification risk bounds. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–1391–III–1399. JMLR.org, 2013.

[PHD20] Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[PL20a] Tomaso Poggio and Qianli Liao. Explicit regularization and implicit bias in deep network classifiers trained with the square loss. *arXiv preprint arXiv:2101.00072*, 2020.

[PL20b] Tomaso Poggio and Qianli Liao. Explicit regularization and implicit bias in deep network classifiers trained with the square loss. *arXiv preprint arXiv:2101.00072*, 2020.

[PS16] Bernardo Ávila Pires and Csaba Szepesvári. Multiclass classification calibration functions. *arXiv preprint arXiv:1609.06385*, 2016.

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[Rif02] Ryan Michael Rifkin. *Everything old is new again: a fresh look at historical approaches in machine learning*. PhD thesis, MaSSachuSettS InStitute of Technology, 2002.

[RK04] Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

[RV+13] Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[SAH19] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[SAH20] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The performance analysis of generalized margin maximizers on separable data. In *International Conference on Machine Learning*, pages 8417–8426. PMLR, 2020.

[SAS22] Vignesh Subramanian, Rahul Arya, and Anant Sahai. Generalization for multiclass classification with overparameterized linear models. In *Advances in Neural Information Processing Systems*, 2022.

[SC19] Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[SFBL98] Robert E Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

[SHN+18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[TB07] Ambuj Tewari and Peter L Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(36):1007–1025, 2007.

[TB20] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

[TOS20] Christos Thrampoulidis, Samet Oymak, and Mahdi Soltanolkotabi. Theoretical insights into multiclass classification: A high-dimensional asymptotic view. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8907–8920. Curran Associates, Inc., 2020.

[TPT20] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Sharp asymptotics and optimal performance for inference in binary models. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3739–3749. PMLR, 26–28 Aug 2020.

[TPT21] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Fundamental limits of ridge-regularized empirical risk minimization in high dimensions. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2773–2781. PMLR, 13–15 Apr 2021.

[Tro06] Joel A Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory*, 52(3):1030–1051, 2006.

[Vap13] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[Wel74] Lloyd Welch. Lower bounds on the maximum cross correlation of signals (corresp.). *IEEE Transactions on Information theory*, 20(3):397–399, 1974.

[WS23] David X Wu and Anant Sahai. Precise asymptotic generalization for multiclass classification with overparameterized linear models. *arXiv preprint arXiv:2306.13255*, 2023.

[WT21] Ke Wang and Christos Thrampoulidis. Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting and regularization. *arXiv preprint arXiv:2011.09148*, 2021.

[WW98] Jason Weston and Chris Watkins. Multi-class support vector machines. Technical report, Citeseer, 1998.

[ZBH+17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[ZDZ+21] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34, 2021.

[Zha04] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Contents

A Lemmas used in the proof of Theorem 2

A.1 Auxiliary Lemmas

In this section, we state a series of auxiliary lemmas that we use to prove Lemma 2. The following result shows concentration of the norms of the label indicators $\mathbf{v}_c, c \in [k]$ under the nearly equal-priors assumption (Assumption 1). Intuitively, in this nearly balanced setting there are $\Theta(n/k)$ samples for each class; hence, $\Theta(n/k)$ non-zeros (in fact, 1's) in each label indicator vector $\mathbf{v}_c$.

Lemma 6. *Under the setting of Assumption 1, there exist large constants $C_1, C_2 > 0$ such that the event*

$$\mathcal{E}_v := \Big\{ \Big(1 - \frac{1}{C_1}\Big)\frac{n}{k} \le \|\mathbf{v}_c\|_2^2 \le \Big(1 + \frac{1}{C_1}\Big)\frac{n}{k} \ , \ \forall c \in [k] \Big\}, \tag{50}$$

holds with probability at least $1 - 2ke^{-\frac{n}{C_2 k^2}}$.

Next, we provide bounds on the "base case" 0-th order quadratic forms that involve the Gram matrix $\mathbf{A}_0^{-1}$. We do this in three lemmas presented below. The first Lemma 7 follows by a direct application of [WT21, Lemma 4 and 5]. The only difference is that we keep track of throughout the proof is the scaling of $\mathcal{O}(1/k)$ arising from the multiclass case in the $\mathbf{v}_j$'s. For instance, the bound of the term $h_{mj}^{(0)} := \mathbf{v}_m^T \mathbf{A}_0^{-1} \mathbf{d}_j$ involves a term $\tilde{\rho}_{n,k} = \min\{1, \sqrt{\log(2n)/k}\}$ compared to the binary case. The other two Lemmas 8 and 9 are proved in Section A.3.

Lemma 7 (0-th order Quadratic forms, Part I)**.** *Under the event $\mathcal{E}_v$, there exist constants c_i's and C_i's > 1 such that the following bounds hold with probability at least $1 - c_1 k e^{-\frac{n}{c_2}}$.*

$$\begin{aligned}
t_{jj}^{(0)} &\le \frac{C_1 n \|\boldsymbol{\mu}\|_2^2}{p} \quad \text{for all } j \in [k], \\
|h_{mj}^{(0)}| &\le \tilde{\rho}_{n,k} \frac{C_2 n \|\boldsymbol{\mu}\|_2}{\sqrt{k} p} \quad \text{for all } m, j \in [k], \\
|t_{mj}^{(0)}| &\le \frac{C_3 n \|\boldsymbol{\mu}\|_2^2}{p} \quad \text{for all } m \ne j \in [k], \\
\|\mathbf{d}_j\|_2^2 &\le C_4 n \|\boldsymbol{\mu}\|_2^2 \quad \text{for all } j \in [k], \\
\max_{i \in [n]} |f_{ji}^{(0)}| &\le \frac{C_5 \sqrt{\log(2n)} \|\boldsymbol{\mu}\|_2}{p} \quad \text{for all } j \in [k].
\end{aligned}$$

To sharply characterize the forms $s_{ij}^{(0)}$ we need additional work, particularly for the cross-terms where $i \ne j$. We will make use of fundamental concentration inequalities on quadratic forms of inverse Wishart matrices. Note that the term $t_{jj}^{(0)}$ originally depends on the norm $\|\boldsymbol{\mu}_j\|_2^2$. Due to the nearly equal energy Assumption 1, we can write $\|\boldsymbol{\mu}_j\|_2^2$ in term of the "reference vector" norm $\|\boldsymbol{\mu}\|_2^2$ (which is defined in Assumption 1). Consequently, we will see this "reference norm" $\|\boldsymbol{\mu}\|_2^2$ in all our higher order terms. The following lemma controls these quadratic forms, and shows in particular that the $s_{ij}^{(0)}$ terms for $i \ne j$ are much smaller than the corresponding terms $s_{jj}^{(0)}$. This sharp control of the cross-terms is essential for several subsequent proof steps.

Lemma 8 (0-th order Quadratic forms, Part II)**.** *Working on the event $\mathcal{E}_v$ defined in Equation (50), assume that $p > Cn\log(kn) + n - 1$ for large enough constant $C > 1$ and large n. There exist constants C_i's > 1 such that with probability at least $1 - \frac{C_0}{n}$, the following bound holds:*

$$\begin{aligned}
\frac{C_1 - 1}{C_1} \cdot \frac{n}{kp} \le & s_{jj}^{(0)} \le \frac{C_1 + 1}{C_1} \cdot \frac{n}{kp}, \quad \text{for } j \in [k], \\
-\frac{C_2 + 1}{C_2} \cdot \frac{\sqrt{n}}{kp} \le & s_{ij}^{(0)} \le \frac{C_2 + 1}{C_2} \cdot \frac{\sqrt{n}}{kp}, \quad \text{for } i \ne j \in [k].
\end{aligned}$$

The proof of Lemma 8 for the cross terms with $i \neq j$ critically uses the in-built orthogonality of the label indicator vectors $\{\mathbf{v}_c\}_{c\in[k]}$. Finally, the following lemma controls the quadratic forms $g_{ji}^{(0)}$.

Lemma 9 (0-th order Quadratic forms, Part III)**.** *Working on the event $\mathcal{E}_v$ defined in Equation (50), given $p > Ck^3 n\log(kn) + n - 1$ for a large constant C, there exist large enough constants C_1, C_2, such that with probability at least $1 - \frac{2}{kn}$, we have for every $i \in [n]$:*

$$\begin{aligned}\left(1 - \frac{1}{C_1}\right)\frac{1}{p} \leq & g_{(y_i)i}^{(0)} \leq \left(1 + \frac{1}{C_1}\right)\frac{1}{p},\\ -\frac{1}{C_2}\cdot\frac{1}{k^2 p} \leq & g_{ji}^{(0)} \leq \frac{1}{C_2}\cdot\frac{1}{k^2 p}, \quad \text{for } j \neq y_i.\end{aligned}$$

A.2 Proof of Lemma 2

In this section, we provide the full proof of Lemma 2. We begin with a proof outline.

A.2.1 Proof outline

As explained in Section 6.2, it suffices to consider the case where $j = k$, since when $j \neq k$ we can simply change the order of adding mean components, described in Equation (39), so that the j-th mean component is added last. For concreteness, we will also fix $i \in [n]$, $y_i = k$ and define as shorthand $m := k - 1$. These fixes are without loss of generality. The reason why we fix $j = k$ and $m = k - 1$ is that when we do the proof, we want to add the $k-1$-th and k-th components last. This is for ease of reading and understanding.

For the case $j = k$, the leave-one-out quadratic forms in Lemma 2 are equal to the quadratic forms of order $k - 1$, given by $s_{kk}^{(k-1)}, t_{kk}^{(k-1)}$, $h_{kk}^{(k-1)}$, $g_{ki}^{(k-1)}$ and $f_{ki}^{(k-1)}$. We will proceed recursively starting from the quadratic forms of order 1 building up all the way to the quadratic forms of order $k - 1$. Specifically, starting from order 1, we will work on the event

$$\mathcal{E}_q := \{\text{all the inequalities in Lemmas 7, 8 and 9 hold}\}, \tag{51}$$

Further, we note that Lemma 9 shows that the bound for $g_{y_i i}^{(0)}$ is different from the bound for $g_{ji}^{(0)}$ when $j \neq y_i$. We will show the following set of upper and lower bounds:

$$\begin{aligned}\left(\frac{C_{11}-1}{C_{11}}\right)\frac{n}{kp} \leq s_{kk}^{(1)} \leq \left(\frac{C_{11}+1}{C_{11}}\right)\frac{n}{kp},\\ -\left(\frac{C_{12}+1}{C_{12}}\right)\frac{\sqrt{n}}{kp} \leq s_{mk}^{(1)} \leq \left(\frac{C_{12}+1}{C_{12}}\right)\frac{\sqrt{n}}{kp},\\ t_{kk}^{(1)} \leq \frac{C_{13} n\|\boldsymbol{\mu}\|_2^2}{p},\\ |h_{mk}^{(1)}| \leq \tilde{\rho}_{n,k}\frac{C_{14} n\|\boldsymbol{\mu}\|_2}{\sqrt{kp}},\\ |t_{mk}^{(1)}| \leq \frac{C_{15} n\|\boldsymbol{\mu}\|_2^2}{p},\\ \|\mathbf{d}_k\|_2^2 \leq C_{16} n\|\boldsymbol{\mu}\|_2^2,\\ |f_{ki}^{(1)}| \leq \frac{C_{17}\sqrt{n}\|\boldsymbol{\mu}\|_2}{p},\\ \left(1 - \frac{1}{C_{18}}\right)\frac{1}{p} \leq g_{(y_i)i}^{(1)} \leq \left(1 + \frac{1}{C_{18}}\right)\frac{1}{p}, \text{ and}\\ -\frac{1}{C_{19}k^2 p} \leq g_{mi}^{(1)} \leq \frac{1}{C_{19}k^2 p}\end{aligned} \tag{52}$$

with probability at least $1 - \frac{c}{kn^2}$. Comparing the bounds on the terms of order 1 in Equation (52) with the terms in Lemmas 7, 8 and 9 of order 0, the key observation is that they are all at the same order.

This allows us to repeat the same argument to now bound corresponding terms of order 2, and so on until order $k-1$. Note that for each $j \in [k]$, we have n terms of the form $g_{ji}^{(1)}$, corresponding to each value of $i \in [n]$. Thus, we will adjust the final probabilities by applying a union bound over the n training examples.

A.2.2 Proofs for 1-st order quadratic forms in Equation (52)

The proof makes repeated use of Lemmas 7, 8 and 9. In fact, we will throughout condition on the event $\mathcal{E}_q$, defined in Equation (51), which holds with probability at least $1 - \frac{c_1}{n} - c_2 e^{-\frac{n}{c_0 k^2}}$. Specifically, by Lemma 7 we have

$$h_{mj}^{(0)} \le \tilde{\rho}_{n,k} \frac{C_1 \epsilon_n}{k^2 \sqrt{n}}, \qquad \max_{i \in [n]} |f_{mi}^{(0)}| \le \frac{C_2 \epsilon_n}{k^{1.5} n}, \quad \text{and} \quad \frac{s_{mj}^{(0)}}{s_{kk}^{(0)}} \le \frac{C}{\sqrt{n}} \text{ for } m, j \neq k, \tag{53}$$

where we recall from Equation (44) the notation $\epsilon_n := \frac{k^{1.5} n \sqrt{n} \|\boldsymbol{\mu}\|_2}{p}$. Also, recall that we choose $\epsilon_n \le \tau$ for a sufficiently small constant τ.

In order to make use of Lemmas 7, 8 and 9, we need to relate the quantities of interest to corresponding quadratic forms involving $\mathbf{A}_0$. We do this recursively and make repeated use of the Woodbury identity. The recursions are proved in Appendix F.1. We now provide the proofs for the bounds on the terms in Equation (52) one-by-one.

Bounds on $s_{mk}^{(1)}$. By Equation (102) in Appendix F.1, we have

$$s_{mk}^{(1)} = s_{mk}^{(0)} - \frac{1}{\det_0} (\star)_s^{(0)}, \tag{54}$$

where we define

$$\begin{aligned} (\star)_s^{(0)} &:= (\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}) s_{1k}^{(0)} s_{1m}^{(0)} + s_{1m}^{(0)} h_{k1}^{(0)} h_{11}^{(0)} + s_{1k}^{(0)} h_{m1}^{(0)} h_{11}^{(0)} - s_{11}^{(0)} h_{k1}^{(0)} h_{m1}^{(0)} + s_{1m}^{(0)} h_{k1}^{(0)} + s_{1k}^{(0)} h_{m1}^{(0)} \text{ and} \\ \det_0 &:= s_{11}^{(0)} (\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}) + (1 + h_{11}^{(0)})^2. \end{aligned} \tag{55}$$

The essential idea is to show that $|\frac{(\star)_s^{(0)}}{\det_0}|$ is sufficiently small compared to $|s_{mk}^{(0)}|$. We first look at the first term given by $\left((\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}) s_{1k}^{(0)} s_{1m}^{(0)}\right) / \det_0$. By Lemmas 7, 8 and the definition of $\det_0$, we have

$$\left| \frac{1}{\det_0} \left((\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}) s_{1k}^{(0)} s_{1m}^{(0)} \right) \right| \le \frac{(\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}) |s_{1k}^{(0)} s_{1m}^{(0)}|}{s_{11}^{(0)} (\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)})} = \left| \frac{s_{1k}^{(0)} s_{1m}^{(0)}}{s_{11}^{(0)}} \right| \le \frac{C_1}{\sqrt{n}} \cdot \frac{C_2 + 1}{C_2} \cdot \frac{\sqrt{n}}{kp},$$

where we use $\det_0 \ge s_{11}^{(0)} (\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)})$ and $s_{mj}^{(0)} / s_{kk}^{(0)} \le C/\sqrt{n}$ for all $m, j \neq k$. Now, we upper bound the other two dominant terms $|s_{1m}^{(0)} h_{k1}^{(0)} / \det_0|$ and $|s_{1k}^{(0)} h_{m1}^{(0)} / \det_0|$. Note that the same bound will apply to the remaining terms in Equation (55) because we trivially have $|h_{ij}^{(0)}| = \mathcal{O}(1)$ for all $(i, j) \in [k]$. Again, Lemmas 7 and 8 give us

$$\left| \frac{s_{1m}^{(0)} h_{k1}^{(0)}}{\det_0} \right| \le \frac{|s_{1m}^{(0)} h_{k1}^{(0)}|}{(1 + h_{11}^{(0)})^2} \le \frac{\tilde{\rho}_{n,k} C_3 \epsilon_n}{\left(1 - \frac{C_5 \tilde{\rho}_{n,k} \epsilon_n}{k^2 \sqrt{n}}\right)^2 k^2 \sqrt{n}} \cdot \frac{C_2 + 1}{C_2} \cdot \frac{\sqrt{n}}{kp}.$$

The identical bound holds for $|s_{1k}^{(0)} h_{m1}^{(0)}|$. Noting that $|s_{mk}^{(0)}| \le \frac{C_2+1}{C_2} \cdot \frac{\sqrt{n}}{kp}$, we then have

$$\begin{aligned} |s_{mk}^{(1)}| &\le |s_{mk}^{(0)}| + \left| \frac{(\star)_s^{(0)}}{\det_0} \right| \\ &\le \left(1 + \frac{C_6}{\sqrt{n}} + \frac{C_7 \tilde{\rho}_{n,k} \epsilon_n}{\left(1 - \frac{C_5 \tilde{\rho}_{n,k} \epsilon_n}{k^2 \sqrt{n}}\right)^2 k^2 \sqrt{n}} \right) \frac{C_2 + 1}{C_2} \cdot \frac{\sqrt{n}}{kp} \\ &\le (1 + \alpha) \cdot \frac{C_2 + 1}{C_2} \cdot \frac{\sqrt{n}}{kp}, \end{aligned} \tag{56}$$

where in the last inequality, we use that $\epsilon \le \tau$ for sufficiently small constant $\tau > 0$, and defined

$$\alpha := \frac{C_6}{\sqrt{n}} + \frac{C_7\tau}{\left(1 - \frac{C_5\tau}{k^2\sqrt{n}}\right)^2 k^2\sqrt{n}}.$$

Now, we pick τ to be sufficiently small and n to be sufficiently large such that $(1+\alpha)\frac{C_2+1}{C_2} \le \frac{C_8+1}{C_8}$ for some constant $C_8 > 0$. Then, we conclude with the following upper bound:

$$|s^{(1)}_{mk}| \le \frac{C_8+1}{C_8} \cdot \frac{\sqrt{n}}{kp}.$$

Bounds on $s^{(1)}_{kk}$. Equation (103) in Appendix F.1 gives us

$$s^{(1)}_{kk} = s^{(0)}_{kk} - \frac{1}{\det_0}\Big((\|\boldsymbol{\mu}_1\|_2^2 - t^{(0)}_{11}){s^{(0)}_{1k}}^2 + 2s^{(0)}_{1k}h^{(0)}_{k1}h^{(0)}_{11} - s^{(0)}_{11}{h^{(0)}_{k1}}^2 + 2s^{(0)}_{1k}h^{(0)}_{k1}\Big).$$

First, we lower bound $s^{(1)}_{kk}$ by upper bounding $\frac{1}{\det_0}\Big((\|\boldsymbol{\mu}_1\|_2^2 - t^{(0)}_{11}){s^{(0)}_{1k}}^2\Big)$. Lemmas 7 and 8 yield

$$\frac{1}{\det_0}\Big((\|\boldsymbol{\mu}_1\|_2^2 - t^{(0)}_{11}){s^{(0)}_{1k}}^2\Big) \le \frac{(\|\boldsymbol{\mu}_1\|_2^2 - t^{(0)}_{11}){s^{(0)}_{1k}}^2}{s^{(0)}_{11}(\|\boldsymbol{\mu}_1\|_2^2 - t^{(0)}_{11}) + (1+h^{(0)}_{11})^2} \le \frac{(\|\boldsymbol{\mu}_1\|_2^2 - t^{(0)}_{11}){s^{(0)}_{1k}}^2}{s^{(0)}_{11}(\|\boldsymbol{\mu}_1\|_2^2 - t^{(0)}_{11})} \le \frac{C_1}{n} \cdot \frac{n}{kp}.$$

It suffices to upper bound the other dominant term $|s^{(0)}_{1k}h^{(0)}_{k1}|/\det_0$. For this term, we have

$$\Big|\frac{s^{(0)}_{1k}h^{(0)}_{k1}}{\det_0}\Big| \le \frac{|s^{(0)}_{1k}h^{(0)}_{k1}|}{(1+h^{(0)}_{11})^2} \le \frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{\left(1 - \frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^2\sqrt{n}} \cdot \frac{C_2+1}{C_2} \cdot \frac{\sqrt{n}}{kp}.$$

Thus, we get

$$s^{(1)}_{kk} \ge \left(1 - \frac{C_1}{n} - \frac{C_5\tilde{\rho}_{n,k}n\epsilon_n}{\left(1 - \frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^2\sqrt{n}}\right)\frac{C_6-1}{C_6} \cdot \frac{n}{kp} \ge (1-\alpha) \cdot \frac{C_6-1}{C_6} \cdot \frac{n}{kp}.$$

Next, we upper bound $s^{(1)}_{kk}$ by a similar argument, and get

$$\begin{aligned} s^{(1)}_{kk} &\le |s^{(0)}_{kk}| + \frac{1}{\det_0}\Big|2s^{(0)}_{1k}h^{(0)}_{k1}h^{(0)}_{11} + s^{(0)}_{11}{h^{(0)}_{k1}}^2 + 2s^{(0)}_{1k}h^{(0)}_{k1}\Big| \\ &\le \left(1 + \frac{C_7\tilde{\rho}_{n,k}\epsilon_n}{\left(1 - \frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^2\sqrt{n}}\right)\frac{C_8+1}{C_8} \cdot \frac{n}{kp} \le (1+\alpha')\frac{C_8+1}{C_8} \cdot \frac{n}{kp}, \end{aligned}$$

where we used $\frac{1}{\det_0}\Big((\|\boldsymbol{\mu}_1\|_2^2 - t^{(0)}_{11}){s^{(0)}_{1k}}^2\Big) > 0$ in the first step. As above, we can tune ϵ and n such that $(1+\alpha')\frac{C_8+1}{C_8} \le \frac{C_9+1}{C_9}$ and $(1-\alpha)\frac{C_6-1}{C_6} \ge \frac{C_9-1}{C_9}$ for sufficiently large constant $C_9 > 0$.

Bounds on $h^{(1)}_{mk}$. Equation (104) in Appendix F.1 gives us

$$h^{(1)}_{mk} = h^{(0)}_{mk} - \frac{1}{\det_0}(\star)^{(0)}_h,$$

where we define

$$(\star)^{(0)}_h = (\|\boldsymbol{\mu}_1\|_2^2 - t^{(0)}_{11})s^{(0)}_{1m}h^{(0)}_{1k} + h^{(0)}_{m1}h^{(0)}_{1k}h^{(0)}_{11} + h^{(0)}_{m1}h^{(0)}_{1k} + s^{(0)}_{1m}t^{(0)}_{k1} + s^{(0)}_{1m}t^{(0)}_{k1}h^{(0)}_{11} - s^{(0)}_{11}t^{(0)}_{k1}h^{(0)}_{m1}.$$

We focus on the two dominant terms $((\|\boldsymbol{\mu}_1\|_2^2 - t^{(0)}_{11})s^{(0)}_{1m}h^{(0)}_{1k})/\det_0$ and $s^{(0)}_{1m}t^{(0)}_{k1}/\det_0$. For the first dominant term $((\|\boldsymbol{\mu}_1\|_2^2 - t^{(0)}_{11})s^{(0)}_{1m}h^{(0)}_{1k})/\det_0$, Lemmas 7 and 8 yield

$$\Big|\frac{1}{\det_0}\Big((\|\boldsymbol{\mu}_1\|_2^2 - t^{(0)}_{11})s^{(0)}_{1m}h^{(0)}_{1k}\Big)\Big| \le \frac{(\|\boldsymbol{\mu}_1\|_2^2 - t^{(0)}_{11})|s^{(0)}_{1m}h^{(0)}_{1k}|}{s^{(0)}_{11}(\|\boldsymbol{\mu}_1\|_2^2 - t^{(0)}_{11})} \le \Big|\frac{s^{(0)}_{1m}h^{(0)}_{1k}}{s^{(0)}_{11}}\Big| \le \frac{C_1}{\sqrt{n}}|h^{(0)}_{1k}| \le \frac{C_2\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}.$$

For the second dominant term $s_{1m}^{(0)} t_{k1}^{(0)} / \det_0$, we have

$$\frac{1}{\det_0} s_{1m}^{(0)} t_{k1}^{(0)} \leq \frac{|s_{1m}^{(0)} t_{k1}^{(0)}|}{(1+h_{11}^{(0)})^2} \leq \frac{C_3 n\sqrt{n}\|\boldsymbol{\mu}\|_2^2}{\left(1-\frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 kp^2} \leq \frac{C_5\epsilon_n}{\left(1-\frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^{1.5}\sqrt{n}} \cdot \frac{\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}},$$

where we use the fact $1/\sqrt{k} < \tilde{\rho}_{n,k}$ for $k > 1$. Thus, we get

$$\begin{aligned} |h_{mk}^{(1)}| &\leq |h_{mk}^{(0)}| + \left|\frac{1}{\det_0}(\star)_h^{(0)}\right| \leq \left(1 + \frac{C_1}{\sqrt{n}} + \frac{C_5\tilde{\rho}_{n,k}\epsilon_n}{\left(1-\frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^{1.5}\sqrt{n}}\right)\frac{C_6\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}} \\ &\leq (1+\alpha)\frac{C_7\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}, \end{aligned}$$

and there exists constant C_8 such that $(1+\alpha)C_7 \leq C_8$, which shows the desired upper bound.

Bounds on $t_{kk}^{(1)}$. Equation (106) in Appendix F.1 gives us

$$t_{kk}^{(1)} = t_{kk}^{(0)} - \frac{1}{\det_0}\left(\left(\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}\right) {h_{1k}^{(0)}}^2 + 2t_{1k}^{(0)} h_{1k}^{(0)} h_{11}^{(0)} - s_{11}^{(0)} {t_{1k}^{(0)}}^2 + 2t_{1k}^{(0)} h_{1k}^{(0)}\right).$$

We only need an upper bound on $t_{kk}^{(1)}$. The first dominant term $s_{11}^{(0)} {t_{1k}^{(0)}}^2 / \det_0$ is upper bounded as follows:

$$\frac{s_{11}^{(0)} {t_{1k}^{(0)}}^2}{\det_0} \leq \frac{s_{11}^{(0)} {t_{1k}^{(0)}}^2}{(1+h_{11}^{(0)})^2} \leq \frac{C_6 n^3\|\boldsymbol{\mu}\|_2^4}{\left(1-\frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 kp^3} \leq \frac{C_7\epsilon_n^2}{\left(1-\frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 pk^4 n} \cdot \frac{n\|\boldsymbol{\mu}\|_2^2}{p}.$$

Next, the second dominant term, $t_{1k}^{(0)} h_{1k}^{(0)} / \det_0$, is upper bounded as

$$\frac{t_{1k}^{(0)} h_{1k}^{(0)}}{\det_0} \leq \frac{|t_{1k}^{(0)} h_{1k}^{(0)}|}{(1+h_{11}^{(0)})^2} \leq \frac{C_8\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}\left(1-\frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2} \cdot \frac{n\|\boldsymbol{\mu}\|_2^2}{p}.$$

Combining the results above gives us

$$\begin{aligned} t_{kk}^{(1)} &\leq t_{kk}^{(0)} + \frac{1}{\det_0}\left|2t_{1k}^{(0)} h_{1k}^{(0)} h_{11}^{(0)} + s_{11}^{(0)} {t_{1k}^{(0)}}^2 + 2t_{1k}^{(0)} h_{1k}^{(0)}\right| \\ &\leq \left(1 + \frac{C_9\tilde{\rho}_{n,k}\epsilon_n}{\left(1-\frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^2\sqrt{n}}\right)\frac{n\|\boldsymbol{\mu}\|_2^2}{p} \leq \frac{C_5 n\|\boldsymbol{\mu}\|_2^2}{p}. \end{aligned}$$

This shows the desired upper bound.

Bounds on $t_{mk}^{(1)}$. Equation (105) in Appendix F.1 gives us

$$t_{mk}^{(1)} = t_{mk}^{(0)} - \frac{1}{\det_0}(\star)_t^{(0)},$$

where we define

$$(\star)_t^{(0)} = (\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}) h_{1m}^{(0)} h_{1k}^{(0)} + t_{m1}^{(0)} h_{1k}^{(0)} h_{11}^{(0)} + t_{k1}^{(0)} h_{1m}^{(0)} h_{11}^{(0)} + t_{1m}^{(0)} h_{1k}^{(0)} + t_{1k}^{(0)} h_{1m}^{(0)} - s_{11}^{(0)} t_{1m}^{(0)} t_{1k}^{(0)}.$$

Again, we only need an upper bound on $t_{mk}^{(1)}$. As in the previously derived bounds, we have

$$\frac{1}{\det_0}(\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}) h_{1m}^{(0)} h_{1k}^{(0)} \leq \frac{(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})|h_{1m}^{(0)} h_{1k}^{(0)}|}{s_{11}^{(0)}(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})} \leq \frac{C_1\tilde{\rho}_{n,k}^2 n^2\|\boldsymbol{\mu}\|_2^2}{kp^2} \cdot \frac{kp}{n} \leq \frac{C_1 n\|\boldsymbol{\mu}\|_2^2}{p}.$$

The other dominant term $t_{1m}^{(0)}h_{1m}^{(0)}/\det_0$ is upper bounded as:

$$\frac{t_{1m}^{(0)}h_{1m}^{(0)}}{\det_0} \leq \frac{|t_{1m}^{(0)}h_{1m}^{(0)}|}{(1+h_{11}^{(0)})^2} \leq \frac{C_2\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}(\left(1-\frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2} \cdot \frac{n\|\boldsymbol{\mu}\|_2^2}{p}.$$

Combining the results above yields

$$\begin{aligned}|t_{mk}^{(1)}| &\leq |t_{mk}^{(0)}| + \frac{1}{\det_0}\left|(\star)_t^{(0)}\right| \\ &\leq \left(C_1 + \frac{C_2\tilde{\rho}_{n,k}\epsilon_n}{\left(1-\frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^2\sqrt{n}}\right)\frac{n\|\boldsymbol{\mu}\|_2^2}{p} \leq \frac{C_4 n\|\boldsymbol{\mu}\|_2^2}{p}.\end{aligned}$$

Note that both $t_{kk}^{(0)}$ and $t_{mk}^{(0)}$ are much smaller than $\|\boldsymbol{\mu}\|_2^2$. The above upper bound shows that this continues to hold for $t_{kk}^{(1)}$ and $t_{mk}^{(1)}$ since $p \gg n$.

Bounds on $f_{ki}^{(1)}$. Consider $i \in [n]$ and fix $y_i = k$ without loss of generality. Equation (107) in Appendix F.1 gives us

$$f_{ki}^{(1)} = f_{ki}^{(0)} - \frac{1}{\det_0}(\star)_f^{(0)}, \tag{57}$$

where we define

$$(\star)_f^{(0)} = (\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)})h_{1k}^{(0)}g_{1i}^{(0)} + t_{1k}^{(0)}g_{1i}^{(0)} + t_{1k}^{(0)}h_{11}^{(0)}g_{1i}^{(0)} + h_{1k}^{(0)}f_{1i}^{(0)} + h_{1k}^{(0)}h_{11}^{(0)}f_{1i}^{(0)} - s_{11}^{(0)}t_{1k}^{(0)}f_{1i}^{(0)}. \tag{58}$$

We only need an upper bound on $f_{ki}^{(1)}$. We consider the dominant terms $(\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)})h_{1k}^{(0)}g_{1i}^{(0)}/\det_0$, $t_{1k}^{(0)}g_{1i}^{(0)}/\det_0$, $h_{1k}^{(0)}f_{1i}^{(0)}/\det_0$ and $s_{11}^{(0)}t_{1k}^{(0)}f_{1i}^{(0)}/\det_0$. Lemmas 7, 8 and 9 give us

$$\begin{aligned}&\frac{(\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)})h_{1k}^{(0)}g_{1i}^{(0)}}{\det_0} \leq \frac{(\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)})|h_{1k}^{(0)}g_{1i}^{(0)}|}{(\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)})s_{11}^{(0)}} \\ &\leq \frac{C_1\tilde{\rho}_{n,k}n\|\boldsymbol{\mu}\|_2}{\sqrt{kp}} \cdot \frac{1}{C_2k^2p} \cdot \frac{kp}{n} \leq \frac{C_3}{k^{1.5}\sqrt{n}} \cdot \frac{\sqrt{n}\|\boldsymbol{\mu}\|_2}{p}, \\ &\frac{t_{1k}^{(0)}g_{1i}^{(0)}}{\det_0} \leq \frac{|t_{1k}^{(0)}g_{1i}^{(0)}|}{(1+h_{11}^{(0)})^2} \leq \frac{C_4 n\|\boldsymbol{\mu}\|_2^2}{\left(1-\frac{C_5\tilde{\rho}_{n,k}\epsilon}{k^2\sqrt{n}}\right)^2 k^2p^2} \leq \frac{C_7\epsilon_n}{\left(1-\frac{C_5\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^{3.5}n} \cdot \frac{\sqrt{n}\|\boldsymbol{\mu}\|_2}{p}, \\ &\frac{h_{1k}^{(0)}f_{1i}^{(0)}}{\det_0} \leq \frac{|h_{1k}^{(0)}f_{1i}^{(0)}|}{(1+h_{11}^{(0)})^2} \leq \frac{C_6\tilde{\rho}_{n,k}\epsilon_n}{\left(1-\frac{C_5\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^2\sqrt{n}} \cdot \frac{\sqrt{n}\|\boldsymbol{\mu}\|_2}{p}, \text{ and} \\ &\frac{s_{11}^{(0)}t_{1k}^{(0)}f_{1i}^{(0)}}{\det_0} \leq \frac{|s_{11}^{(0)}t_{1k}^{(0)}f_{1i}^{(0)}|}{(1+h_{11}^{(0)})^2} \leq \frac{C_7\epsilon_n^2}{k^4n\left(1-\frac{C_5\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2} \cdot \frac{\sqrt{n}\|\boldsymbol{\mu}\|_2}{p},\end{aligned}$$

where, in the last two steps, we used the upper bound $C\sqrt{n}\|\boldsymbol{\mu}\|_2/p$ for $|f_{ji}^{(0)}|$ and previously derived bounds on $|h_{1k}^{(0)}|$ and $|s_{11}^{(0)}t_{1k}^{(0)}|$. Thus, we have

$$\begin{aligned}|f_{ki}^{(1)}| &\leq |f_{ki}^{(0)}| + \left|\frac{1}{\det_0}(\star)_f^{(0)}\right| \\ &\leq \left(1 + \frac{C_3}{k^{1.5}\sqrt{n}} + \frac{C_8\epsilon_n}{\left(1-\frac{C_5\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^2\sqrt{n}}\right)\frac{C_9\sqrt{n}\|\boldsymbol{\mu}\|_2}{p} \\ &\leq (1+\alpha)\frac{C_{10}\epsilon_n}{k^{1.5}n},\end{aligned}$$

and we have $(1+\alpha)C_{10} \le C_{11}$ for a large enough positive constant C_{11}. This shows the desired upper bound.

Bounds on $g_{ki}^{(1)}$ and $g_{mi}^{(1)}$. Equation (108) in Appendix F.1 gives

$$z_{ci}\mathbf{e}_i^T\mathbf{A}_1^{-1}\mathbf{u}_k = |z_{ci}|^2\Big(\mathbf{e}_i^T\mathbf{A}_0^{-1}\mathbf{v}_k - \frac{1}{\det_0}(\star)_{gk}^{(0)}\Big) = |z_{ci}|^2\Big(g_{ki}^{(0)} - \frac{1}{\det_0}(\star)_{gk}^{(0)}\Big), \tag{59}$$

where we define

$$(\star)_{gk}^{(0)} = (\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)})s_{1k}^{(0)}g_{1i}^{(0)} + g_{1i}^{(0)}h_{11}^{(0)}h_{k1}^{(0)} + g_{1i}^{(0)}h_{k1}^{(0)} + s_{1k}^{(0)}f_{1i}^{(0)} + s_{1k}^{(0)}h_{11}^{(0)}f_{1i}^{(0)} - s_{11}^{(0)}h_{k1}^{(0)}f_{1i}^{(0)}.$$

Lemmas 7, 8 and 9 give us

$$\begin{aligned}
&\frac{(\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)})|s_{1k}^{(0)}g_{1i}^{(0)}|}{\det_0} \le \frac{(\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)})|s_{1k}^{(0)}g_{1i}^{(0)}|}{(\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)})s_{11}^{(0)}} \le \frac{C_1}{\sqrt{n}}\cdot\frac{1}{C_2k^2p},\\
&\frac{|h_{k1}^{(0)}g_{1i}^{(0)}|}{\det_0} \le \frac{|h_{k1}^{(0)}g_{1i}^{(0)}|}{(1+h_{11}^{(0)})^2} \le \frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{\Big(1-\frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\Big)^2k^2\sqrt{n}}\cdot\frac{1}{C_2k^2p}, \text{ and}\\
&\frac{|s_{1k}^{(0)}f_{1i}^{(0)}|}{\det_0} \le \frac{|s_{1k}^{(0)}f_{1i}^{(0)}|}{(1+h_{11}^{(0)})^2} \le \frac{C_5\epsilon_n}{\Big(1-\frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\Big)^2\sqrt{k}\sqrt{n}}\cdot\frac{1}{C_2k^2p}.
\end{aligned}$$

We then have

$$\begin{aligned}
g_{ki}^{(1)} &\ge g_{ki}^{(0)} - \frac{1}{\det_0}|(\star)_{gk}^{(0)}| \ge \Big(1-\frac{1}{C}\Big)\left(1 - \frac{C_1}{k^2\sqrt{n}} - \frac{C_6\epsilon_n}{\Big(1-\frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\Big)^2k^{2.5}\sqrt{n}}\right)\frac{1}{p} \ge \Big(1-\frac{1}{C}\Big)\frac{1-\alpha}{p}\\
g_{ki}^{(1)} &\le g_{ki}^{(0)} + \frac{1}{\det_0}|(\star)_{gk}^{(0)}| \le \Big(1+\frac{1}{C}\Big)\left(1 + \frac{C_1}{k^2\sqrt{n}} + \frac{C_7\epsilon_n}{\Big(1-\frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\Big)^2k^{2.5}\sqrt{n}}\right)\frac{1}{p} \le \Big(1+\frac{1}{C}\Big)\frac{1+\alpha}{p},
\end{aligned}$$

where for large enough n and positive constant C_9, we have $(1+\alpha)\frac{C+1}{C} \le \frac{C_9+1}{C_9}$ and $(1-\alpha)\frac{C-1}{C} \ge \frac{C_9-1}{C_9}$. Similarly, for the case $m \ne k$, we have

$$z_{ci}\mathbf{e}_i^T\mathbf{A}_1^{-1}\mathbf{u}_m = |z_{ci}|^2\left(\mathbf{e}_i^T\mathbf{A}_0^{-1}\mathbf{v}_m - \frac{1}{\det_0}(\star)_{gm}^{(0)}\right) = |z_{ci}|^2\Big(g_{mi}^{(0)} - \frac{1}{\det_0}(\star)_{gm}^{(0)}\Big), \tag{60}$$

where we define

$$(\star)_{gm}^{(0)} = (\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)})s_{1m}^{(0)}g_{1i}^{(0)} + g_{1i}^{(0)}h_{11}^{(0)}h_{m1}^{(0)} + g_{1i}^{(0)}h_{m1}^{(0)} + s_{1m}^{(0)}f_{1i}^{(0)} + s_{1m}^{(0)}h_{11}^{(0)}f_{1i}^{(0)} - s_{11}^{(0)}h_{m1}^{(0)}f_{1i}^{(0)}.$$

As a consequence of our nearly equal energy and priors assumption (Assumption 1), we can directly use the bounds of the terms in $(\star)_{gk}^{(0)}$ to bound terms in $(\star)_{gm}^{(0)}$. We get

$$|g_{mi}^{(1)}| \le \frac{1}{C}\left(1 + \frac{C_1}{\sqrt{n}} + \frac{C_8\epsilon_n}{(1-(\frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}))^2\sqrt{k}\sqrt{n}}\right)\frac{1}{k^2p} \le \frac{1}{C}\cdot\frac{1+\alpha}{k^2p}.$$

Finally, there exists a sufficiently large constant C_{10} such that $(1+\alpha)/C \le 1/C_{10}$. This shows the desired bounds.

A.2.3 Completing the proof for k-th order quadratic forms

Notice from the above analysis that the 1-st order quadratic forms exhibit the same order-wise dependence on n, k and p as the 0-th order quadratic forms, e.g. both $s_{mk}^{(0)}$ and $s_{mk}^{(1)}$ are of order $\Theta(\frac{\sqrt{n}}{kp})$. Thus, the higher-order quadratic forms that arise by including more mean components will not

change too much[6]. By Equation (39), we can see that we can bound the 2-nd order quadratic forms by bounding quadratic forms with order 1. We consider $s^{(2)}_{mk}$ as an example:

$$s^{(2)}_{mk} = s^{(1)}_{mk} - \frac{1}{\det_1}(\star)^{(1)}_s,$$

where

$$(\star)^{(1)}_s := (\|\boldsymbol{\mu}_2\|_2^2 - t^{(1)}_{22})s^{(1)}_{2k}s^{(1)}_{2m} + s^{(1)}_{2m}h^{(1)}_{k2}h^{(1)}_{22} + s^{(1)}_{2k}h^{(1)}_{m2}h^{(1)}_{22} - s^{(1)}_{22}h^{(1)}_{k2}h^{(1)}_{m2} + s^{(1)}_{2m}h^{(1)}_{k2} + s^{(1)}_{2k}h^{(1)}_{m2},$$
$$\det_1 := s^{(1)}_{22}(\|\boldsymbol{\mu}_2\|_2^2 - t^{(1)}_{22}) + (1 + h^{(1)}_{22})^2.$$

We additionally show how $f^{(2)}_{ki}$ relates to the 1-st order quadratic forms:

$$f^{(2)}_{ki} = f^{(1)}_{ki} - \frac{1}{\det_1}(\star)^{(1)}_f,$$

where we define

$$(\star)^{(1)}_f = (\|\boldsymbol{\mu}_2\|_2^2 - t^{(1)}_{22})h^{(1)}_{2k}g^{(1)}_{2i} + t^{(1)}_{2k}g^{(1)}_{2i} + t^{(1)}_{2k}h^{(1)}_{22}g^{(1)}_{2i} + h^{(1)}_{2k}f^{(1)}_{2i} + h^{(1)}_{2k}h^{(1)}_{22}f^{(1)}_{2i} - s^{(1)}_{22}t^{(1)}_{2k}f^{(1)}_{2i}.$$

Observe that the equations above are very similar to Equations (54) and (55) (for s), and Equations (57) and (58) (for f), except that the quadratic forms are in terms of Gram matrix $\mathbf{A}_1$. We have shown that the quadratic forms with order 1 will not be drastically different different from the quadratic forms with order 0. Hence, we repeat the above procedures of bounding these quadratic forms $k-1$ times to obtain the desired bounds in Lemma 2. The only quantity that will change in each iteration is α, which nevertheless remains negligible[7].

Our analysis so far is conditioned on event $\mathcal{E}_q$. We define the *unconditional* event $\mathcal{E}_u := \{$all the inequalities in Lemma 2 hold$\}$. Then, we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_u^c) &\le \mathbb{P}(\mathcal{E}_u^c|\mathcal{E}_q) + \mathbb{P}(\mathcal{E}_q^c) \le \mathbb{P}(\mathcal{E}_u^c|\mathcal{E}_q) + \mathbb{P}(\mathcal{E}_q^c|\mathcal{E}_v) + \mathbb{P}(\mathcal{E}_v^c)\\ &\le \frac{c_1}{kn} + \frac{c_2}{n} + c_3k(e^{-\frac{n}{c_4}} + e^{-\frac{n}{c_5k^2}})\\ &\le \frac{c_6}{n} + c_7ke^{-\frac{n}{c_5k^2}},\end{aligned}$$

for constants c_i's > 1. This completes the proof.

A.3 Proofs of Auxiliary lemmas

We complete this section by proving the auxiliary Lemmas 6, 8 and 9, which were used in the proof of Lemma 2.

A.3.1 Proof of Lemma 6

Our goal is to upper and lower bound $\|\mathbf{v}_c\|_2^2$, for $c \in [k]$. Note that every entry of $\mathbf{v}_c$ is either 1 or 0, hence these entries are independent sub-Gaussian random variables with sub-Gaussian parameter 1 [Wai19, Chapter 2]. Recall that under the nearly equal-prior Assumption 1, we have $(1-(1/C_1))(n/k) \le \mathbb{E}[\|\mathbf{v}_c\|_2^2] \le (1+(1/C_2))(n/k)$ for large enough constants $C_1, C_2 > 0$. Thus, a straightforward application of Hoeffding's concentration inequality on bounded random variables [Wai19, Chapter 2] gives us

$$\mathbb{P}\left(\left|\|\mathbf{v}_c\|_2^2 - \mathbb{E}[\|\mathbf{v}_c\|_2^2]\right| \ge t\right) \le 2\exp\left(-\frac{t^2}{2n}\right).$$

We complete the proof by setting $t = \frac{n}{C_3k}$ for a large enough constant C_3 and applying the union bound over all $c \in [k]$.

[6]There are several low-level reasons for this. One critical reason is the aforementioned orthogonality of the label indicator vectors $\{\mathbf{v}_c\}_{c\in[k]}$, which ensures by Lemma 8 that the cross-terms $|s^{(j)}_{mk}|$ are always dominated by the larger terms $|s^{(j)}_{kk}|$. Another reason is that $h^{(0)}_{mk}$, which can be seen as the "noise" term in our analysis, is small and thus does not affect other terms.

[7]To see this, recall that in the first iteration we had $\alpha_1 := \alpha = \frac{C_1}{\sqrt{n}} + \frac{C_2\tau}{(1-(C_5\tau/(k^2\sqrt{n})))^2k^2\sqrt{n}}$ for the first-order terms. Thus, even if we repeat the procedure $k-1$ times, then we have $\alpha_k \le Ck\alpha_1$, which remains small since we consider $n \gg k$.

A.3.2 Proof of Lemma 8

We use the following lemma adapted from [MNS+21, Lemma 2] to bound quadratic forms of inverse Wishart matrices.

Lemma 10. *Define $p'(n) := (p-n+1)$, and consider matrix $\mathbf{M} \sim$ Wishart$(p, \mathbf{I}_n)$. For any unit Euclidean norm vector $\mathbf{v}$ and any $t > 0$, we have*

$$\mathbb{P}\Big(\frac{1}{\mathbf{v}^T\mathbf{M}^{-1}\mathbf{v}} > p'(n) + \sqrt{2tp'(n)} + 2t\Big) \le e^{-t} \quad \text{and} \quad \mathbb{P}\Big(\frac{1}{\mathbf{v}^T\mathbf{M}^{-1}\mathbf{v}} < p'(n) - \sqrt{2tp'(n)}\Big) \le e^{-t},$$

provided that $p'(n) > 2\max\{t, 1\}$.

We first upper and lower bound $s_{cc}^{(0)}$ for a fixed $c \in [k]$. Recall that we assume $p > Cn\log(kn) + n - 1$ for sufficiently large constant $C > 1$ and this can be obtained by assuming $p'(n) > Cn\log(kn)$. Let $t = 2\log(kn)$. Working on the event $\mathcal{E}_v$ defined in (50), Lemma 10 gives us

$$s_{cc}^{(0)} \le \frac{\|\mathbf{v}_c\|_2^2}{p'(n) - \sqrt{4\log(kn)p'(n)}} \le \frac{C_1+1}{C_1} \cdot \frac{n/k}{p'(n)\left(1 - \frac{2}{\sqrt{Cn}}\right)} \le \frac{C_2+1}{C_2} \cdot \frac{n}{kp}$$

with probability at least $1 - \frac{2}{k^2n^2}$. Here, the last inequality comes from the fact that p is sufficiently large compared to n and C is large enough. Similarly, for the lower bound, we have

$$s_{cc}^{(0)} \ge \frac{\|\mathbf{v}_c\|_2^2}{p'(n) + \sqrt{4\log(kn)p'(n)} + 2\log(kn)} \ge \frac{C_1-1}{C_1} \cdot \frac{n/k}{p'(n)\left(1 + \frac{4}{\sqrt{Cn}}\right)} \ge \frac{C_2-1}{C_2} \cdot \frac{n}{kp}$$

with probability $1 - \frac{2}{k^2n^2}$.

Now we upper and lower bound $s_{cj}^{(0)}$ for a fixed choice $j \ne c \in [k]$. We use the parallelogram law to get

$$\mathbf{v}_c^T\mathbf{A}_0^{-1}\mathbf{v}_j = \frac{1}{4}\Big((\mathbf{v}_c+\mathbf{v}_j)^T\mathbf{A}_0^{-1}(\mathbf{v}_c+\mathbf{v}_j) - (\mathbf{v}_c-\mathbf{v}_j)^T\mathbf{A}_0^{-1}(\mathbf{v}_c-\mathbf{v}_j)\Big).$$

Because of the orthogonality of the label indicator vectors ($\mathbf{v}_c^T\mathbf{v}_j = 0$ for any $j \ne c$), we have $\|\mathbf{v}_c+\mathbf{v}_j\|_2^2 = \|\mathbf{v}_c-\mathbf{v}_j\|_2^2$, which we denote by $\tilde{n}$ as shorthand. Then, we have

$$\begin{aligned}
\mathbf{v}_c^T\mathbf{A}_0^{-1}\mathbf{v}_j &\le \frac{1}{4}\left(\frac{\tilde{n}}{p'(n) - \sqrt{4\log(kn)p'(n)}} - \frac{\tilde{n}}{p'(n) + \sqrt{4\log(kn)p'(n)} + 4\log(kn)}\right) \\
&\le \frac{1}{4} \cdot \frac{2\tilde{n}\sqrt{4\log(kn)p'(n)} + 4\tilde{n}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})(p'(n) + \sqrt{4\log(kn)p'(n)})} \\
&\le \frac{C_1+1}{2C_1k} \cdot \frac{2n\sqrt{4\log(kn)p'(n)} + 4n\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})(p'(n) + \sqrt{4\log(kn)p'(n)})}
\end{aligned}$$

with probability at least $1 - \frac{2}{k^2n^2}$ Here, the last inequality follows because we have $\tilde{n} \le \frac{2(C_1+1)}{C_1} \cdot \frac{n}{k}$ on $\mathcal{E}_v$. Because $p'(n) > Cn\log(kn)$, we have

$$\begin{aligned}
\mathbf{v}_c^T\mathbf{A}_0^{-1}\mathbf{v}_j &\le \frac{C_1+1}{2C_1k} \cdot \frac{2\sqrt{n}p'(n) \cdot \sqrt{4/C} + 4/C \cdot p'(n)}{\left(1 - \sqrt{4/(Cn)}\right)p'(n)^2} \\
&\le \frac{C_1+1}{2C_1} \cdot \frac{\sqrt{n}}{k} \cdot \frac{2\sqrt{4/C} + \sqrt{4/(Cn)}}{p'(n)(1 - \sqrt{4/(Cn)})} \\
&\le \frac{C_2+1}{C_2} \cdot \frac{\sqrt{n}}{kp},
\end{aligned}$$

where in the last step we use the fact that $C > 1$ is large enough. To lower bound $s_{cj}^{(0)}$, we get

$$\begin{aligned}
\mathbf{v}_c^T \mathbf{A}_0^{-1} \mathbf{v}_j &\geq \frac{1}{4}\left(\frac{\tilde{n}}{(p'(n) + \sqrt{4\log(kn)p'(n)} + 4\log(kn))} - \frac{\tilde{n}}{(p'(n) - \sqrt{4\log(kn)p'(n)})}\right) \\
&\geq \frac{1}{4} \cdot \frac{-2\tilde{n}\sqrt{4\log(kn)p'(n)} - 4\tilde{n}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})(p'(n) + \sqrt{4\log(kn)p'(n)})} \\
&\geq -\frac{C_1+1}{2C_1 k} \cdot \frac{2n\sqrt{4\log(kn)p'(n)} + 4n\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})(p'(n) + \sqrt{4\log(kn)p'(n)})}
\end{aligned}$$

with probability at least $1 - \frac{2}{k^2n^2}$. Then following similar steps to the upper bound of $\mathbf{v}_c^T \mathbf{A}_0^{-1} \mathbf{v}_j$ gives us

$$\begin{aligned}
\mathbf{v}_c^T \mathbf{A}_0^{-1} \mathbf{v}_j &\geq -\frac{C_1+1}{2C_1 k} \cdot \frac{2\sqrt{n}p'(n)\sqrt{4/C} + (4/C)p'(n)}{(p'(n) - \sqrt{4/(Cn)}p'(n))p'(n)} \\
&\geq -\frac{C_1+1}{2C_1} \cdot \frac{\sqrt{n}}{k} \cdot \frac{2\sqrt{4/C} + (4/C\sqrt{n})}{p'(n)(1 - \sqrt{4/(Cn)})} \\
&\geq -\frac{C_2+1}{C_2} \cdot \frac{\sqrt{n}}{kp}.
\end{aligned}$$

We finally apply the union bound on all pairs of $c, j \in [k]$ and complete the proof.

A.3.3 Proof of Lemma 9

We first lower and upper bound $g_{(y_i)i}^{(0)}$. Recall that we assumed $y_i = k$ without loss of generality. With a little abuse of notation, we define $\|\mathbf{v}_k\|_2^2 = \tilde{n}$ and $\mathbf{u} := \sqrt{\tilde{n}}\mathbf{e}_i$. We use the parallelogram law to get

$$\mathbf{e}_i^T \mathbf{A}_0^{-1} \mathbf{v}_k = \frac{1}{4\sqrt{\tilde{n}}}\left((\mathbf{u} + \mathbf{v}_k)^T \mathbf{A}_0^{-1}(\mathbf{u} + \mathbf{v}_k) - (\mathbf{u} - \mathbf{v}_k)^T \mathbf{A}_0^{-1}(\mathbf{u} - \mathbf{v}_k)\right).$$

Note that $\|\mathbf{u} + \mathbf{v}_k\|_2^2 = 2(\tilde{n} + \sqrt{\tilde{n}})$ and $\|\mathbf{u} - \mathbf{v}_k\|_2^2 = 2(\tilde{n} - \sqrt{\tilde{n}})$. As before, we apply Lemma 10 with $t = 2\log(kn)$ to get with probability at least $1 - \frac{2}{k^2n^2}$,

$$\begin{aligned}
\mathbf{e}_i^T \mathbf{A}_0^{-1} \mathbf{v}_k &\geq \frac{1}{4\sqrt{\tilde{n}}}\left(\frac{2(\tilde{n} + \sqrt{\tilde{n}})}{(p'(n) + \sqrt{4\log(kn)p'(n)} + 4\log(kn))} - \frac{2(\tilde{n} - \sqrt{\tilde{n}})}{(p'(n) - \sqrt{4\log(kn)p'(n)})}\right) \\
&\geq \frac{1}{4\sqrt{\tilde{n}}} \cdot \frac{4\sqrt{\tilde{n}}p'(n) - 4\tilde{n}\sqrt{4\log(kn)p'(n)} - 8\tilde{n}\log(kn)}{(p'(n) + \sqrt{4\log(kn)p'(n)} + 4\log(kn))p'(n)} \\
&\geq \frac{p'(n) - \sqrt{\tilde{n}}\sqrt{4\log(kn)p'(n)} - 2\sqrt{\tilde{n}}\log(kn)}{(p'(n) + \sqrt{4\log(kn)p'(n)} + 4\log(kn))p'(n)}, \\
&\geq \frac{p'(n) - \sqrt{(1+1/C_1)n/k}\sqrt{4\log(kn)p'(n)} - 2\sqrt{(1+1/C_1)n/k}\log(kn)}{(p'(n) + \sqrt{4\log(kn)p'(n)} + 4\log(kn))p'(n)}.
\end{aligned}$$

The last inequality works on event $\mathcal{E}_v$, by which we have $\tilde{n} \leq \frac{2(C_1+1)n}{C_1 k}$. Then, $p'(n) > Ck^3 n\log(kn)$ gives us

$$\begin{aligned}
\mathbf{e}_i^T \mathbf{A}_0^{-1} \mathbf{v}_k &\geq \frac{p'(n) - \sqrt{(1+1/C_1)n/k}\sqrt{4/(Ck^3n)}p'(n) - \sqrt{(1+1/C_1)n/k}(2/Ck^3n)p'(n)}{(p'(n) + \sqrt{4\log(kn)p'(n)} + 4\log(kn))p'(n)} \\
&\geq \frac{1 - (1/(C_2\sqrt{k^4})) - (1/(C_3 k^{3.5}\sqrt{n}))}{p'(n)(1 + 2\sqrt{4/(Ck^3n)})} \\
&\geq \frac{C_4 - 1}{C_4} \cdot \frac{1}{p},
\end{aligned}$$

where in the last step we use the fact that $C, C_2, C_3 > 1$ are large enough. To upper bound $g^{(0)}_{(y_i)i}$, we have with probability at least $1 - \frac{2}{k^2n^2}$,

$$\begin{aligned}
\mathbf{e}_i^T \mathbf{A}_0^{-1} \mathbf{v}_k &\le \frac{1}{4\sqrt{\tilde{n}}} \left(\frac{2(\tilde{n} + \sqrt{\tilde{n}})}{(p'(n) - \sqrt{4\log(kn)p'(n)})} - \frac{2(\tilde{n} - \sqrt{\tilde{n}})}{(p'(n) + \sqrt{4\log(kn)p'(n)} + 4\log(kn))} \right) \\
&\le \frac{1}{4\sqrt{\tilde{n}}} \cdot \frac{4\sqrt{\tilde{n}}p'(n) + 4\tilde{n}\sqrt{4\log(kn)p'(n)} + 8\tilde{n}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)} \\
&\le \frac{p'(n) + \sqrt{\tilde{n}}\sqrt{4\log(kn)p'(n)} + 2\sqrt{\tilde{n}}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)}, \\
&\le \frac{p'(n) + \sqrt{(1+1/C_1)n/k}\sqrt{4\log(kn)p'(n)} + 2\sqrt{(1+1/C_1)n/k}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)}.
\end{aligned}$$

Then $p'(n) > Ck^3 n\log(kn)$ gives us

$$\begin{aligned}
\mathbf{e}_i^T \mathbf{A}_0^{-1} \mathbf{v}_k &\le \frac{p'(n) + \sqrt{(1+1/C_1)n/k}\sqrt{4/(Ck^3n)}p'(n) + 2\sqrt{(1+1/C_1)n/k}(4/Ck^3n)p'(n)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)} \\
&\le \frac{1 + (1/(C_2\sqrt{k^4})) + (1/(C_3k^{3.5}\sqrt{n}))}{p'(n)(1 - 2\sqrt{4/(Ck^3n)})} \\
&\le \frac{C_4 + 1}{C_4} \cdot \frac{1}{p}.
\end{aligned}$$

We now upper and lower bound $g^{(0)}_{ji}$ for a fixed $j \neq y_i$. As before, we have

$$\mathbf{e}_i^T \mathbf{A}_0^{-1} \mathbf{v}_j = \frac{1}{4\sqrt{\tilde{n}}} \left((\mathbf{u} + \mathbf{v}_j)^T \mathbf{A}_0^{-1} (\mathbf{u} + \mathbf{v}_j) - (\mathbf{u} - \mathbf{v}_j)^T \mathbf{A}_0^{-1} (\mathbf{u} - \mathbf{v}_j) \right).$$

Since $\mathbf{e}_i^T \mathbf{v}_j = 0$, we now have $\|\mathbf{u} + \mathbf{v}_j\|_2^2 = \|\mathbf{u} - \mathbf{v}_j\|_2^2 = 2\tilde{n}$. We apply Lemma 10 with $t = 2\log(kn)$ to get, with probability at least $1 - \frac{2}{k^2n^2}$,

$$\begin{aligned}
\mathbf{e}_i^T \mathbf{A}_0^{-1} \mathbf{v}_j &\le \frac{1}{4\sqrt{\tilde{n}}} \left(\frac{2\tilde{n}}{(p'(n) - \sqrt{4\log(kn)p'(n)})} - \frac{2\tilde{n}}{(p'(n) + \sqrt{4\log(kn)p'(n)} + 4\log(kn))} \right) \\
&\le \frac{1}{4\sqrt{\tilde{n}}} \cdot \frac{4\tilde{n}\sqrt{4\log(kn)p'(n)} + 8\tilde{n}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)} \\
&\le \frac{\sqrt{\tilde{n}}\sqrt{4\log(kn)p'(n)} + 2\sqrt{\tilde{n}}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)}, \\
&\le \frac{\sqrt{(1+1/C_1)n/k}\sqrt{4\log(kn)p'(n)} + 2\sqrt{(1+1/C_1)n/k}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)}.
\end{aligned}$$

The last inequality works on event $\mathcal{E}_v$, by which we have $\tilde{n} \le \frac{2(C_1+1)n}{C_1 k}$. Then, $p'(n) > Ck^3 n\log(kn)$ gives us

$$\begin{aligned}
\mathbf{e}_i^T \mathbf{A}_0^{-1} \mathbf{v}_j &\le \frac{\sqrt{(1+1/C_1)n/k}\sqrt{4/(Ck^3n)}p'(n) + \sqrt{(1+1/C_1)n/k}(2/Ck^3n)p'(n)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)} \\
&\le \frac{(1/(C_2\sqrt{k^4})) + (1/(C_3k^{3.5}\sqrt{n}))}{p'(n)(1 - \sqrt{4/(Ck^3n)})} \\
&\le \frac{C_4 + 1}{C_4} \cdot \frac{1}{k^2 p},
\end{aligned}$$

where in the last step we use the fact that $C, C_2, C_3 > 1$ are large enough. To lower bound $g_{ij}^{(0)}$, we have with probability at least $1 - \frac{2}{k^2 n^2}$,

$$\begin{aligned}
\mathbf{e}_i^T \mathbf{A}_0^{-1} \mathbf{v}_j &\geq \frac{1}{4\sqrt{\tilde{n}}} \left(\frac{2\tilde{n}}{(p'(n) + \sqrt{4\log(kn)p'(n)} + 4\log(kn))} - \frac{2\tilde{n}}{(p'(n) - \sqrt{4\log(kn)p'(n)})} \right) \\
&\geq \frac{1}{4\sqrt{\tilde{n}}} \cdot \frac{-4\tilde{n}\sqrt{4\log(kn)p'(n)} - 8\tilde{n}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)} \\
&\geq -\frac{\sqrt{\tilde{n}}\sqrt{4\log(kn)p'(n)} + 2\sqrt{\tilde{n}}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)}, \\
&\geq -\frac{\sqrt{(1+1/C_1)n/k}\sqrt{4\log(kn)p'(n)} + 2\sqrt{(1+1/C_1)n/k}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)}.
\end{aligned}$$

Because $p'(n) > Ck^3 n \log(kn)$, we get

$$\begin{aligned}
\mathbf{e}_i^T \mathbf{A}_0^{-1} \mathbf{v}_j &\geq -\frac{\sqrt{(1+1/C_1)n/k}\sqrt{4/(Ck^3 n)}p'(n) + \sqrt{(1+1/C_1)n/k}(2/Ck^3 n)p'(n)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)} \\
&\geq -\frac{(1/(C_2\sqrt{k^4})) + (1/(C_3 k^{3.5}\sqrt{n}))}{p'(n)(1 - \sqrt{4/(Ck^3 n)})} \\
&\geq -\frac{C_4 + 1}{C_4} \cdot \frac{1}{k^2 p},
\end{aligned}$$

where in the last step we use the fact that $C, C_2, C_3 > 1$ are large enough. We complete the proof by applying a union bounds over all k classes and n training examples.

B Proof of Theorem 3

In this section, we provide the proof of Theorem 3, which was discussed in Section 3.2.2. After having derived the interpolation condition in Equation (12) for multiclass SVM, the proofs is in fact a rather simple extension of the arguments provided in [MNS+21, HMX21] to the multiclass case. This is unlike the GMM case that we considered in Section 6.2, which required substantial additional effort over and above the binary case [WT21].

For this section, we define $\mathbf{A} = \mathbf{X}^T\mathbf{X}$ as shorthand (we denoted the same quantity as $\mathbf{A}_k$ in Section 6.2). Recall that the eigendecomposition of the covariance matrix is given by $\boldsymbol{\Sigma} = \sum_{i=1}^p \lambda_i \mathbf{v}_i \mathbf{v}_i^T = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$. By rotation invariance of the standard normal variable, we can write $\mathbf{A} = \mathbf{Q}^T\boldsymbol{\Lambda}\mathbf{Q}$, where the entries of $\mathbf{Q} \in \mathbb{R}^{p\times n}$ are IID $\mathcal{N}(0,1)$ random variables. Finally, recall that we denoted $\boldsymbol{\lambda} = \begin{bmatrix}\lambda_1 & \cdots & \lambda_p\end{bmatrix}$ and defined the effective dimensions $d_2 = \frac{\|\boldsymbol{\lambda}\|_1^2}{\|\boldsymbol{\lambda}\|_2^2}$ and $d_\infty = \frac{\|\boldsymbol{\lambda}\|_1}{\|\boldsymbol{\lambda}\|_\infty}$. Observe that Equation (12) in Theorem 1 is equivalent to the condition

$$z_{ci}\mathbf{e}_i^T \mathbf{A}^{-1}\mathbf{z}_c > 0, \quad \text{for all} \quad c \in [k] \quad \text{and} \quad i \in [n]. \tag{61}$$

We fix $c \in [k]$ and drop the subscript c, using $\overline{\mathbf{z}}$ to denote the vector $\mathbf{z}_c$. We first provide a deterministic equivalence to Equation (12) that resembles the condition provided in [HMX21, Lemma 1]. Our proof is slightly modified compared to [HMX21, Lemma 1] and relies on elementary use of block matrix inversion identity.

Lemma 11. *Let $\mathbf{Q} \in \mathbb{R}^{p\times n} = [\mathbf{q}_1, \cdots, \mathbf{q}_n]$. In our notation, Equation (12) holds for a fixed c if and only if:*

$$\frac{1}{z_i}\overline{\mathbf{z}}_{\setminus i}^T \left(\mathbf{Q}_{\setminus i}^T \boldsymbol{\Lambda} \mathbf{Q}_{\setminus i}\right)^{-1} \mathbf{Q}_{\setminus i}^T \boldsymbol{\Lambda} \mathbf{q}_i < 1, \quad \textit{for all} \quad i = 1, \cdots, n. \tag{62}$$

Above, $\overline{\mathbf{z}}_{\setminus i} \in \mathbb{R}^{(n-1)\times 1}$ is obtained by removing the i-th entry from vector $\overline{\mathbf{z}}$ and $\mathbf{Q}_{\setminus i} \in \mathbb{R}^{d\times(n-1)}$ is obtained by removing the i-th column from $\mathbf{Q}$.

Proof. By symmetry, it suffices to consider the case $i = 1$. We first write

$$\mathbf{A} = \begin{bmatrix} \mathbf{q}_1^T \mathbf{\Lambda} \mathbf{q}_1 & \mathbf{q}_1^T \mathbf{\Lambda} \mathbf{Q}_{\setminus 1} \\ \mathbf{Q}_{\setminus 1}^T \mathbf{\Lambda} \mathbf{q}_1 & \mathbf{Q}_{\setminus 1}^T \mathbf{\Lambda} \mathbf{Q}_{\setminus 1} \end{bmatrix} \triangleq \begin{bmatrix} \alpha & \mathbf{b}^T \\ \mathbf{b} & \mathbf{D} \end{bmatrix}.$$

By Schur complement [Ber09], we have

$$\mathbf{A} \succ \mathbf{0} \text{ \textit{iff} either } \left\{ \alpha > 0 \text{ and } \mathbf{D} - \frac{\mathbf{b}\mathbf{b}^T}{\alpha} \succ \mathbf{0} \right\} \text{ or } \left\{ \mathbf{D} \succ \mathbf{0} \text{ and } \alpha - \mathbf{b}^T \mathbf{D}^{-1} \mathbf{b} > 0 \right\}.$$

Since the entries of $\mathbf{Q}$ are drawn from a continuous distribution (IID standard Gaussian), both $\mathbf{A}$ and $\mathbf{D} = \mathbf{Q}_{\setminus 1}^T \mathbf{\Lambda} \mathbf{Q}_{\setminus 1}$ are positive definite almost surely. Therefore, $\alpha - \mathbf{b}^T \mathbf{D}^{-1} \mathbf{b} > 0$ almost surely.

Thus, by block matrix inversion identity [Ber09], we have

$$\mathbf{A}^{-1} = \begin{bmatrix} (\alpha - \mathbf{b}^T \mathbf{D}^{-1} \mathbf{b})^{-1} & -(\alpha - \mathbf{b}^T \mathbf{D}^{-1} \mathbf{b})^{-1} \mathbf{b}^T \mathbf{D}^{-1} \\ -\mathbf{D}^{-1} \mathbf{b} (\alpha - \mathbf{b}^T \mathbf{D}^{-1} \mathbf{b})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1} \mathbf{b} (\alpha - \mathbf{b}^T \mathbf{D}^{-1} \mathbf{b})^{-1} \mathbf{b}^T \mathbf{D}^{-1} \end{bmatrix}.$$

Therefore,$\mathbf{e}_1^T \mathbf{A}^{-1} = (\alpha - \mathbf{b}^T \mathbf{D}^{-1} \mathbf{b})^{-1} \begin{bmatrix} 1 & -\mathbf{b}^T \mathbf{D}^{-1} \end{bmatrix}$. Hence we have

$$z_1 \mathbf{e}_1^T \mathbf{A}^{-1} \overline{\mathbf{z}} = (\alpha - \mathbf{b}^T \mathbf{D}^{-1} \mathbf{b})^{-1} (z_1^2 - \mathbf{b}^T \mathbf{D}^{-1} (z_1 \overline{\mathbf{z}}_{\setminus 1})),$$

where we use the fact that $\overline{\mathbf{z}}_1 = z_1$. Since $\alpha - \mathbf{b}^T \mathbf{D}^{-1} \mathbf{b} > 0$ almost surely, we have

$$\begin{aligned} z_1 \mathbf{e}_1^T \mathbf{A}^{-1} \overline{\mathbf{z}} > 0 &\iff (\alpha - \mathbf{b}^T \mathbf{D}^{-1} \mathbf{b})^{-1} (z_1^2 - \mathbf{b}^T \mathbf{D}^{-1} (z_1 \overline{\mathbf{z}}_{\setminus 1})) > 0 \\ &\iff \frac{1}{z_1} \mathbf{b}^T \mathbf{D}^{-1} \overline{\mathbf{z}}_{\setminus 1} < 1. \end{aligned}$$

Recall that $\mathbf{b}^T = \mathbf{q}_1^T \mathbf{\Lambda} \mathbf{Q}_{\setminus 1}$ and $\mathbf{D} = \mathbf{Q}_{\setminus 1}^T \mathbf{\Lambda} \mathbf{Q}_{\setminus 1}$. This completes the proof. □

Next, we define the following events:

1. For $i \in [n]$, $\mathcal{B}_i := \left\{ \frac{1}{z_i} \overline{\mathbf{z}}_{\setminus i}^T \mathbf{A}_{\setminus i}^{-1} \mathbf{Q}_{\setminus i}^T \mathbf{\Lambda} \mathbf{q}_i \geq 1 \right\}$.
2. For $i \in [n]$, given $t > 0$, $\mathcal{E}_i(t) := \left\{ \| (\overline{\mathbf{z}}_{\setminus i}^T \mathbf{A}_{\setminus i}^{-1} \mathbf{Q}_{\setminus i}^T \mathbf{\Lambda})^T \|_2^2 \geq \frac{1}{t} \right\}$.
3. $\mathcal{B} := \cup_{i=1}^n \mathcal{B}_i$.

We know all the data points are support vectors i.e. Equation (61) holds, if none of the events $\mathcal{B}_i$ happens; hence, $\mathcal{B}$ is the undesired event. We want to upper bound the probability of event $\mathcal{B}$. As in the argument provided in [HMX21], we have

$$\mathbb{P}(\mathcal{B}) \leq \sum_{i=1}^n \Big(\mathbb{P}(\mathcal{B}_i | \mathcal{E}_i(t)^c) + \mathbb{P}(\mathcal{E}_i(t)) \Big). \tag{63}$$

The lemma below gives an upper bound on $\mathbb{P}(\mathcal{B}_i | \mathcal{E}_i(t)^c)$.

Lemma 12. *For any $t > 0$, $\mathbb{P}(\mathcal{B}_i | \mathcal{E}_i(t)^c) \leq 2 \exp\left(-\frac{t}{2ck^2}\right)$.*

Proof. On the event $\mathcal{E}_i(t)^c$, we have $\| (\overline{\mathbf{z}}_{\setminus i}^T \mathbf{A}_{\setminus i}^{-1} \mathbf{Q}_{\setminus i}^T \mathbf{\Lambda})^T \|_2^2 \leq \frac{1}{t}$. Since, by its definition, $|\frac{1}{z_i}| \leq k$, we have $\frac{1}{z_i} \overline{\mathbf{z}}_{\setminus i}^T \mathbf{A}_{\setminus i}^{-1} \mathbf{Q}_{\setminus i}^T \mathbf{\Lambda} \mathbf{q}_i$ is conditionally sub-Gaussian [Wai19, Chapter 2] with parameter at most $ck^2 \| (\overline{\mathbf{z}}_{\setminus i}^T \mathbf{A}_{\setminus i}^{-1} \mathbf{Q}_{\setminus i}^T \mathbf{\Lambda})^T \|_2^2 \leq ck^2/t$. Then the sub-Gaussian tail bound gives

$$\mathbb{P}(\mathcal{B}_i | \mathcal{E}_i(t)^c) \leq 2 \exp\left(-\frac{t}{2ck^2} \right), \tag{64}$$

which completes the prof. □

Next we upper bound $\mathbb{P}(\mathcal{E}_i(t))$ with $t = d_\infty/(2n)$. Since $\|\mathbf{z}_{\setminus i}\|_2 \leq \|\mathbf{y}_{\setminus i}\|_2$, we can directly use [HMX21, Lemma 4].

Lemma 13 (Lemma 4, [HMX21]). $\mathbb{P}\left(\mathcal{E}_i\left(\frac{d_\infty}{2n}\right)\right) \leq 2 \cdot 9^{n-1} \cdot \exp\left(-c_1 \min\left\{\frac{d_2}{4c^2}, \frac{d_\infty}{c}\right\}\right)$.

The results above are proved for fixed choices of $i \in [n]$ and $c \in [k]$. We combine Lemmas 12 and 13 with a union bound over all n training examples and k classes to upper bound the probability of the undesirable event $\mathcal{B}$ over all k classes by:

$$kn9^{n-1} \cdot \exp\left(-c_1 \min\left\{\frac{d_2}{4c^2}, \frac{d_\infty}{c}\right\}\right) \leq \exp\left(-c_1 \min\left\{\frac{d_2}{4c^2}, \frac{d_\infty}{c}\right\} + C_1 \log(kn) + C_2 n\right)$$
$$\text{and } 2kn \cdot \exp\left(-\frac{d_\infty}{2ck^2 n}\right) \leq \exp\left(-\frac{c_2 d_\infty}{ck^2 n} + C_3 \log(kn)\right).$$

Thus, the probability that every data point is a support vector is at least

$$1 - \exp\left(-c_1 \min\left\{\frac{d_2}{4c^2}, \frac{d_\infty}{c}\right\} + C_1 \log(kn) + C_2 n\right) - \exp\left(-\frac{c_2 d_\infty}{ck^2 n} + C_3 \log(kn)\right).$$

To ensure that $\exp\left(-c_1 \min\left\{\frac{d_2}{4c^2}, \frac{d_\infty}{c}\right\} + C_1 \log(kn) + C_2 n\right) + \exp\left(-\frac{c_2 d_\infty}{ck^2 n} + C_3 \log(kn)\right) \leq \frac{c_4}{n}$, we consider the conditions $c_1 \min\left\{\frac{d_2}{4c^2}, \frac{d_\infty}{c}\right\} - C_1 \log(kn) - C_2 n \geq \log(n)$ and $\frac{c_2 d_\infty}{ck^2 n} - C_3 \log(kn) \geq \log(n)$ to be satisfied. These are equivalent to the conditions provided in Equation (17). This completes the proof. Note that throughout the proof, we did not use any generative model assumptions on the labels given the covariates, so in fact our proof applies to scenarios beyond the MLM. □

C Classification error proofs for GMM

In this section, we provide the proofs of classification error under the GMM (Theorem 4 and Theorem 6).

C.1 Proof of Theorem 4

C.1.1 Proof strategy and notations

The notation and main arguments of this proof follow closely the content of Section 6.2.

Our starting point here is the lemma below (adapted from [TOS20, D.10]) that provides a simpler upper bound on the class-wise error $\mathbb{P}_{e|c}$.

Lemma 14. *Under GMM,* $\mathbb{P}_{e|c} \leq \sum_{j \neq c} Q\left(\frac{(\widehat{\mathbf{w}}_c - \widehat{\mathbf{w}}_j)^T \boldsymbol{\mu}_c}{\|\widehat{\mathbf{w}}_c - \widehat{\mathbf{w}}_j\|_2}\right)$. *In particular, if* $(\widehat{\mathbf{w}}_c - \widehat{\mathbf{w}}_j)^T \boldsymbol{\mu}_c > 0$, *then* $\mathbb{P}_{e|c} \leq \sum_{j \neq c} \exp\left(-\frac{((\widehat{\mathbf{w}}_c - \widehat{\mathbf{w}}_j)^T \boldsymbol{\mu}_c)^2}{4(\widehat{\mathbf{w}}_c^T \widehat{\mathbf{w}}_c + \widehat{\mathbf{w}}_j^T \widehat{\mathbf{w}}_j)}\right)$.

Proof. [TOS20, D.10] shows $\mathbb{P}_{e|c}$ is upper bounded by $\sum_{j \neq c} Q\left(\frac{(\widehat{\mathbf{w}}_c - \widehat{\mathbf{w}}_j)^T \boldsymbol{\mu}_c}{\|\widehat{\mathbf{w}}_c - \widehat{\mathbf{w}}_j\|_2}\right)$. Then if $(\widehat{\mathbf{w}}_c - \widehat{\mathbf{w}}_j)^T \boldsymbol{\mu}_c > 0$, the Chernoff bound [Wai19, Ch. 2] gives

$$\mathbb{P}_{e|c} \leq \sum_{j \neq c} \exp\left(-\frac{((\widehat{\mathbf{w}}_c - \widehat{\mathbf{w}}_j)^T \boldsymbol{\mu}_c)^2}{2\|\widehat{\mathbf{w}}_c - \widehat{\mathbf{w}}_j\|_2^2}\right) \leq \sum_{j \neq c} \exp\left(-\frac{((\widehat{\mathbf{w}}_c - \widehat{\mathbf{w}}_j)^T \boldsymbol{\mu}_c)^2}{4(\widehat{\mathbf{w}}_c^T \widehat{\mathbf{w}}_c + \widehat{\mathbf{w}}_j^T \widehat{\mathbf{w}}_j)}\right),$$

where the last inequality uses the identity $\mathbf{a}^T \mathbf{b} \leq 2(\mathbf{a}^T \mathbf{a} + \mathbf{b}^T \mathbf{b})$. □

Thanks to Lemma 14, we can upper bound $P_{e|c}$ by lower bounding the terms

$$\frac{((\widehat{\mathbf{w}}_c - \widehat{\mathbf{w}}_j)^T \boldsymbol{\mu}_c)^2}{(\widehat{\mathbf{w}}_c^T \widehat{\mathbf{w}}_c + \widehat{\mathbf{w}}_j^T \widehat{\mathbf{w}}_j)}, \quad \text{for all} \quad c \neq j \in [k]. \tag{65}$$

Our key observation is that this can be accomplished without the need to control the more intricate cross-correlation terms $\widehat{\mathbf{w}}_c^T \widehat{\mathbf{w}}_j$ for $c \neq j \in [k]$.

Without loss of generality, we assume onwards that $c = k$ and $j = k-1$ (as in Section 6.2). Similar to Section 6.2, the quadratic forms introduced in Equation (40) play key role here, as well. For convenience, we recall the definitions of the *c-th order quadratic forms* for $c, j, m \in [k]$ and $i \in [n]$:

$$\begin{aligned} s_{mj}^{(c)} &:= \mathbf{v}_m^T \mathbf{A}_c^{-1} \mathbf{v}_j, \\ t_{mj}^{(c)} &:= \mathbf{d}_m^T \mathbf{A}_c^{-1} \mathbf{d}_j, \\ h_{mj}^{(c)} &:= \mathbf{v}_m^T \mathbf{A}_c^{-1} \mathbf{d}_j, \\ g_{ji}^{(c)} &:= \mathbf{v}_j^T \mathbf{A}_c^{-1} \mathbf{e}_i, \\ f_{ji}^{(c)} &:= \mathbf{d}_j^T \mathbf{A}_c^{-1} \mathbf{e}_i. \end{aligned}$$

Further, recall that $\widehat{\mathbf{w}}_c = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_c$ and $\mathbf{X} = \sum_{j=1}^k \boldsymbol{\mu}_j \mathbf{v}_j^T + \mathbf{Q}$. Thus,

$$\begin{aligned} \widehat{\mathbf{w}}_c^T \boldsymbol{\mu}_c &= \|\boldsymbol{\mu}_c\|_2^2 \mathbf{v}_c^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_c + \sum_{m \neq c} \boldsymbol{\mu}_m^T \boldsymbol{\mu}_c \mathbf{v}_m^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_c + \mathbf{v}_c^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{d}_c \text{ and} \\ \widehat{\mathbf{w}}_j^T \boldsymbol{\mu}_c &= \|\boldsymbol{\mu}_c\|_2^2 \mathbf{v}_j^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_c + \boldsymbol{\mu}_j^T \boldsymbol{\mu}_c \mathbf{v}_j^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_j + \sum_{m \neq c,j} \boldsymbol{\mu}_m^T \boldsymbol{\mu}_c \mathbf{v}_m^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_j + \mathbf{v}_j^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{d}_c. \end{aligned} \tag{66}$$

Additionally,

$$\widehat{\mathbf{w}}_c^T \widehat{\mathbf{w}}_c = \mathbf{v}_c^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_c, \quad \text{and} \quad \widehat{\mathbf{w}}_j^T \widehat{\mathbf{w}}_j = \mathbf{v}_j^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_j.$$

To lower bound $\widehat{\mathbf{w}}_c^T \boldsymbol{\mu}_c - \widehat{\mathbf{w}}_j^T \boldsymbol{\mu}_c$, we first focus on the dominant terms of Equation (66),

$$\|\boldsymbol{\mu}_c\|_2^2 \mathbf{v}_c^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_c + \mathbf{v}_c^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{d}_c - \|\boldsymbol{\mu}_c\|_2^2 \mathbf{v}_c^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_j - \mathbf{v}_j^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{d}_c. \tag{67}$$

The above terms dominate the bound because, according to Assumption 2, the inner products between different mean vectors are small compared to the norms of mean vectors.

We now lower bound Equation (67) divided by $(\widehat{\mathbf{w}}_c^T \widehat{\mathbf{w}}_c + \widehat{\mathbf{w}}_j^T \widehat{\mathbf{w}}_j)$. Using the leave-one-out trick in Section 6.2 and the matrix-inversion lemma, we show in Appendix C.1.5 that

$$\begin{aligned} &\frac{(67)}{(\widehat{\mathbf{w}}_c^T \widehat{\mathbf{w}}_c + \widehat{\mathbf{w}}_j^T \widehat{\mathbf{w}}_j)} = \frac{D_1}{D_2}, \\ &D_1 = \left(\frac{\|\boldsymbol{\mu}_c\|_2^2 s_{cc}^{(j)} - s_{cc}^{(j)} t_{cc}^{(j)} + {h_{cc}^{(j)}}^2 + h_{cc}^{(j)} - \|\boldsymbol{\mu}_c\|_2^2 s_{jc}^{(j)} - h_{jc}^{(j)} - h_{jc}^{(j)} h_{cc}^{(j)} + s_{jc}^{(j)} t_{cc}^{(j)}}{\det_j} \right)^2, \\ &D_2 = \left(\frac{s_{cc}^{(j)}}{\det_j} + \frac{s_{jj}^{(-j)}}{\det_{-j}} \right), \end{aligned} \tag{68}$$

where $\det_j = (\|\boldsymbol{\mu}_c\|_2^2 - t_{cc}^{(j)}) s_{cc}^{(j)} + (h_{cc}^{(j)} + 1)^2$. Note that $\det_j = \det_{-c}$ when $c = k$ and $j = k-1$.

Next, we will prove that

$$(68) \geq \|\boldsymbol{\mu}\|_2^2 \frac{\left(\left(1 - \frac{C_1}{\sqrt{n}} - \frac{C_2 n}{p} \right) \|\boldsymbol{\mu}\|_2 - C_3 \min\{\sqrt{k}, \sqrt{\log(2n)}\} \right)^2}{C_6 \left(\|\boldsymbol{\mu}\|_2^2 + \frac{kp}{n} \right)}. \tag{69}$$

C.1.2 Proof of Equation (69)

We will lower bound the numerator and upper bound the denominator of Equation (68). We will work on the high-probability event $\mathcal{E}_v$ defined in Equation (50) in Appendix A.1. For quadratic forms such

as $s_{cc}^{(j)}, t_{cc}^{(j)}$ and $h_{cc}^{(j)}$, the Gram matrix $\mathbf{A}_j^{-1}$ does not "include" the c-th mean component because we have fixed $c = k, j = k-1$. Thus, we can directly apply Lemma 2 to get

$$\begin{aligned}\frac{C_1-1}{C_1}\cdot\frac{n}{kp} \leq & s_{cc}^{(j)} \leq \frac{C_1+1}{C_1}\cdot\frac{n}{kp},\\ t_{cc}^{(j)} \leq & \frac{C_2 n\|\boldsymbol{\mu}\|_2^2}{p},\\ -\tilde{\rho}_{n,k}\frac{C_3 n\|\boldsymbol{\mu}\|_2}{\sqrt{k}p} \leq & h_{cc}^{(j)} \leq \tilde{\rho}_{n,k}\frac{C_3 n\|\boldsymbol{\mu}\|_2}{\sqrt{k}p},\end{aligned}$$

on the event $\mathcal{E}_v$. We need some additional work to bound $s_{jc}^{(j)} = \mathbf{v}_j\mathbf{A}_j^{-1}\mathbf{v}_c$ and $h_{jc}^{(j)} = \mathbf{v}_j\mathbf{A}_j^{-1}\mathbf{d}_c$, since the Gram matrix $\mathbf{A}_j^{-1}$ "includes" $\mathbf{v}_j$. The proof here follows the machinery introduced in Appendix A.2 for proving Lemma 2. We provide the core argument and refer the reader therein for additional justifications. By Equation (102) in Appendix F.1 (with the index $j-1$ replacing the index 0), we first have

$$s_{jc}^{(j)} = s_{jc}^{(j-1)} - \frac{1}{\det_{j-1}}(\star)_s^{(j-1)},$$

where we define

$$(\star)_s^{(j-1)} = (\|\boldsymbol{\mu}_j\|_2^2 - t_{jj}^{(j-1)})s_{jj}^{(j-1)}s_{jc}^{(j-1)} + s_{jc}^{(j-1)}h_{jj}^{(j-1)^2} + s_{jc}^{(j-1)}h_{jj}^{(j-1)} + s_{jj}^{(j-1)}h_{jc}^{(j-1)},$$

and $\det_{j-1} = (\|\boldsymbol{\mu}_j\|_2^2 - t_{jj}^{(j-1)})s_{jj}^{(j-1)} + (h_{jj}^{(j-1)}+1)^2$. Further, we have

$$\begin{aligned}|s_{jc}^{(j)}| &= \left|\left(1 - \frac{(\|\boldsymbol{\mu}_j\|_2^2 - t_{jj}^{(j-1)})s_{jj}^{(j-1)} + h_{jj}^{(j-1)^2}}{\det_{j-1}}\right)s_{jc}^{(j-1)} - \frac{1}{\det_{j-1}}(s_{jc}^{(j-1)}h_{jj}^{(j-1)} + s_{jj}^{(j-1)}h_{jc}^{(j-1)})\right|\\ &\leq \frac{1}{C}|s_{jc}^{(j-1)}| + \frac{1}{\det_{j-1}}|(s_{jc}^{(j-1)}h_{jj}^{(j-1)} + s_{jj}^{(j-1)}h_{jc}^{(j-1)})|.\end{aligned}$$

We focus on the dominant term $|s_{jj}^{(j-1)}h_{jc}^{(j-1)}|$. Using a similar argument to that provided in Appendix A.2, we get

$$\begin{aligned}\frac{|s_{jj}^{(j-1)}h_{jc}^{(j-1)}|}{\det_{j-1}} \leq \frac{|s_{jj}^{(j-1)}h_{jc}^{(j-1)}|}{(1+h_{jj}^{(j-1)})^2} &\leq \frac{C_1}{\left(1-\frac{C_2\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2}\cdot\frac{n}{kp}\cdot\frac{\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\\ &\leq \frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{\left(1-\frac{C_2\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^2}\cdot\frac{\sqrt{n}}{kp}.\end{aligned}$$

Thus, we have

$$|s_{jc}^{(j-1)}| \leq \frac{C_4+1}{C_4}\cdot\frac{\sqrt{n}}{kp}.$$

Similarly, we bound the remaining term $h_{jc}^{(j)}$. Specifically, by Equation (104) in Section F.1, we have

$$h_{jc}^{(j)} = h_{jc}^{(j-1)} - \frac{1}{\det_{j-1}}(\star)_h^{(j-1)},$$

where we define

$$(\star)_h^{(j-1)} = (\|\boldsymbol{\mu}_j\|_2^2 - t_{jj}^{(j-1)})s_{jj}^{(j-1)}h_{jc}^{(j-1)} + h_{jc}^{(j-1)}h_{jj}^{(j-1)^2} + h_{jc}^{(j-1)}h_{jj}^{(j-1)} + s_{jj}^{(j-1)}t_{jc}^{(j-1)}.$$

Furthermore,

$$|h_{jc}^{(j)}| = \left| \left(1 - \frac{(\|\boldsymbol{\mu}_j\|_2^2 - t_{jj}^{(j-1)})s_{jj}^{(j-1)} + {h_{jj}^{(j-1)}}^2}{\det_{j-1}}\right) h_{jc}^{(j-1)} - \frac{1}{\det_{j-1}}(h_{jc}^{(j-1)}h_{jj}^{(j-1)} + s_{jj}^{(j-1)}t_{jc}^{(j-1)}) \right|$$
$$\leq \frac{1}{C}|h_{jc}^{(j-1)}| + \frac{1}{\det_{j-1}}|(h_{jc}^{(j-1)}h_{jj}^{(j-1)} + s_{jj}^{(j-1)}t_{jc}^{(j-1)})|.$$

We again consider the dominant term $|s_{jj}^{(j-1)}t_{jc}^{(j-1)}|/\det_{j-1}$ and get

$$\frac{|s_{jj}^{(j-1)}t_{jc}^{(j-1)}|}{\det_{j-1}} \leq \frac{|s_{jj}^{(j-1)}t_{jc}^{(j-1)}|}{(1+h_{jj}^{(j-1)})^2} \leq \frac{C_1}{\left(1 - \frac{C_2\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2} \cdot \frac{n}{kp} \cdot \frac{n\|\boldsymbol{\mu}\|_2^2}{p}$$
$$\leq \frac{C_3\epsilon_n}{\left(1 - \frac{C_2\tilde{\rho}_{n,k}\epsilon_n}{k^{1.5}\sqrt{n}}\right)^2 k^2\sqrt{n}} \cdot \frac{\tilde{\rho}_{n,k}n\|\boldsymbol{\mu}\|_2}{\sqrt{k}p}.$$

Thus, we find that

$$|h_{jc}^{(j-1)}| \leq \tilde{\rho}_{n,k}\frac{C_4 n\|\boldsymbol{\mu}\|_2}{\sqrt{k}p}.$$

We are now ready to lower bound the RHS in Equation (68) by lower bounding its numerator and upper bounding its denominator.

First, for the numerator we have the following sequence of inequalities:

$$\begin{aligned}
&\|\boldsymbol{\mu}_c\|_2^2 s_{cc}^{(j)} - s_{cc}^{(j)}t_{cc}^{(j)} + {h_{cc}^{(j)}}^2 + h_{cc}^{(j)} - \|\boldsymbol{\mu}_c\|_2^2 s_{jc}^{(j)} - h_{jc}^{(j)} - h_{jc}^{(j)}h_{cc}^{(j)} + s_{jc}^{(j)}t_{cc}^{(j)} \\
\geq &\|\boldsymbol{\mu}_c\|_2^2 s_{cc}^{(j)} - \|\boldsymbol{\mu}_c\|_2^2 s_{jc}^{(j)} - s_{cc}^{(j)}t_{cc}^{(j)} + s_{jc}^{(j)}t_{cc}^{(j)} + h_{cc}^{(j)} - h_{jc}^{(j)} - h_{jc}^{(j)}h_{cc}^{(j)} \\
\geq &\frac{C_1 - 1}{C_1} \cdot \frac{\|\boldsymbol{\mu}\|_2^2 n}{kp} - \frac{C_2 + 1}{C_2} \cdot \frac{\|\boldsymbol{\mu}\|_2^2\sqrt{n}}{kp} - \frac{C_3 n}{p} \cdot \frac{\|\boldsymbol{\mu}\|_2^2 n}{kp} - \frac{C_4 n}{p} \cdot \frac{\|\boldsymbol{\mu}\|_2^2\sqrt{n}}{kp} - \frac{C_5\tilde{\rho}_{n,k}n\|\boldsymbol{\mu}\|_2}{\sqrt{k}p}.
\end{aligned}$$

Above, we use the fact that the terms $|h_{cc}^{(j)}|, |h_{jc}^{(j)}| \leq C\epsilon/(k^2\sqrt{n})$ are sufficiently small compared to 1. Consequently, the numerator is lower bounded by

$$\left(\frac{C_1 - 1}{C_1} \cdot \frac{\|\boldsymbol{\mu}\|_2^2 n}{kp} - \frac{C_2 + 1}{C_2} \cdot \frac{\|\boldsymbol{\mu}\|_2^2\sqrt{n}}{kp} - \frac{C_3 n}{p} \cdot \frac{\|\boldsymbol{\mu}\|_2^2 n}{kp} - \frac{C_4 n}{p} \cdot \frac{\|\boldsymbol{\mu}\|_2^2\sqrt{n}}{kp} - \frac{C_5\tilde{\rho}_{n,k}n\|\boldsymbol{\mu}\|_2}{\sqrt{k}p}\right)^2 \Big/ \det_j^2. \tag{70}$$

Second, we upper bound the denominator. For this, note that under the assumption of nearly equal energy and equal priors on class means (Assumption 1), there exist constants $C_1, C_2 > 0$ such that $C_1 \leq \det_j / \det_{-j} \leq C_2$. (In fact, a very similar statement was proved in Equation (45) and used in the proof of Theorem 2). Moreover, Lemma 2 shows that the terms $s_{cc}^{(j)}$ and $s_{jj}^{(-j)}$ are of the same order, so it suffices to upper bound $\frac{s_{cc}^{(j)}}{\det_j}$. Again applying Lemma 2, we have

$$\frac{s_{cc}^{(j)}}{\det_j} \leq \frac{C_6}{\det_j} \cdot \frac{n}{kp} \tag{71}$$

on the event $\mathcal{E}_v$. Then, combining Equations (70) and (71) gives us

$$\begin{aligned}
(68) &\geq \frac{n}{C_0 kp} \cdot \frac{1}{\det_j}\left((1 - \frac{C_1}{\sqrt{n}} - \frac{C_2 n}{p})\|\boldsymbol{\mu}\|_2^2 - C_3\min\{\sqrt{k}, \sqrt{\log(2n)}\}\|\boldsymbol{\mu}\|_2\right)^2 \\
&\geq \frac{n}{C_0 kp} \cdot \frac{1}{\frac{C_4\|\boldsymbol{\mu}\|_2^2 n}{kp} + 2 + \frac{C_5 n^2\|\boldsymbol{\mu}\|_2^2}{kp^2}}\left(\left(1 - \frac{C_1}{\sqrt{n}} - \frac{C_2 n}{p}\right)\|\boldsymbol{\mu}\|_2^2 - C_3\min\{\sqrt{k}, \sqrt{\log(2n)}\}\|\boldsymbol{\mu}\|_2\right)^2 \\
&\geq \|\boldsymbol{\mu}\|_2^2 \frac{\left(\left(1 - \frac{C_1}{\sqrt{n}} - \frac{C_2 n}{p}\right)\|\boldsymbol{\mu}\|_2 - C_3\min\{\sqrt{k}, \sqrt{\log(2n)}\}\right)^2}{C_6\left(\|\boldsymbol{\mu}\|_2^2 + \frac{kp}{n}\right)},
\end{aligned} \tag{72}$$

where the second inequality follows from the following upper bound on $\det_j$ on the event $\mathcal{E}_v$:

$$\det_j = (\|\boldsymbol{\mu}_c\|_2^2 - t_{cc}^{(j)})s_{cc}^{(j)} + (h_{cc}^{(j)} + 1)^2 \leq \|\boldsymbol{\mu}_c\|_2^2 s_{cc}^{(j)} + 2({h_{cc}^{(j)}}^2 + 1) \leq \frac{C_4\|\boldsymbol{\mu}\|_2^2 n}{kp} + 2 + \frac{C_5 n^2\|\boldsymbol{\mu}\|_2^2}{kp^2}.$$

C.1.3 Bounding the remaining terms in (66)

The previous sections of the proof bounded the dominant terms in Equation (66). Now, we turn to bounding the remaining terms $\boldsymbol{\mu}_m^T\boldsymbol{\mu}_c\mathbf{v}_m^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_c$ and $\boldsymbol{\mu}_j^T\boldsymbol{\mu}_c\mathbf{v}_j^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_j$. Under the nearly equal energy and priors assumption, the $\mathbf{v}_j^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_j$ terms have the same bound for every $j \in [k]$ except for some constants. Similarly, the $\mathbf{v}_j^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_m$ terms also have the same bound for all $j \neq m \in [k]$ except for some constants. An upper bound on classification error can then be derived in terms of the inner products between the mean vectors. Specifically, we need to include the bounds of $\sum_{m\neq c}\boldsymbol{\mu}_m^T\boldsymbol{\mu}_c\mathbf{v}_m^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_c$, $\boldsymbol{\mu}_j^T\boldsymbol{\mu}_c\mathbf{v}_j^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_j$ and $\sum_{m\neq c,j}\boldsymbol{\mu}_m^T\boldsymbol{\mu}_c\mathbf{v}_m^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_j$. Recall that in Appendix C.1.4 we show that $\mathbf{v}_c(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_j = (s_{cj}^{(j)} + s_{cj}^{(j)}h_{cc}^{(j)} - s_{cc}^{(j)}h_{jc}^{(j)})/\det_j$ and $\mathbf{v}_j(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_j = s_{jj}^{(-j)}/\det_{-j}$. We also show that the bound for $s_{cc}^{(j)}$ (also $s_{jj}^{(-j)}$) is at the order of $\mathcal{O}(n/(kp))$ and the bound for $s_{cj}^{(j)}$ is at the order of $\mathcal{O}(\sqrt{n}/(kp))$ when $c \neq j$, which is significantly smaller than $\mathcal{O}(n/(kp))$ when n is large. Additionally, the bound for $|h_{jc}^{(j)}|$ is sufficiently small. Combining these results and the assumption of mutually incoherent means, we can see that the bounds for these additional terms included are still much smaller than the bound of $\|\boldsymbol{\mu}\|_2^2\mathbf{v}_c^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_c$, which is the dominant term. Therefore, they will not change the generalization bound of (68) except up to constant factors.

C.1.4 Completing the proof

Because of our assumption of nearly equal energy on class means and equal priors, the analysis above can be applied to bound $\frac{((\widehat{\mathbf{w}}_c-\widehat{\mathbf{w}}_j)^T\boldsymbol{\mu}_c)^2}{(\widehat{\mathbf{w}}_c^T\widehat{\mathbf{w}}_c+\widehat{\mathbf{w}}_j^T\widehat{\mathbf{w}}_j)}$, for every $j \neq c$ and $c \in [k]$. We define the *unconditional* event

$$\mathcal{E}_{u2} := \left\{ \frac{((\widehat{\mathbf{w}}_c - \widehat{\mathbf{w}}_j)^T\boldsymbol{\mu}_c)^2}{(\widehat{\mathbf{w}}_c^T\widehat{\mathbf{w}}_c + \widehat{\mathbf{w}}_j^T\widehat{\mathbf{w}}_j)} \text{ is lower bounded by (72) for every } j \neq c \right\}.$$

We have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{u2}^c) &\leq \mathbb{P}(\mathcal{E}_{u2}^c|\mathcal{E}_v) + \mathbb{P}(\mathcal{E}_v^c) \\ &\leq \frac{c_4}{n} + c_5 k(e^{-\frac{n}{c_6}} + e^{-\frac{n}{c_7k^2}}) \leq \frac{c_4}{n} + c_8 k e^{-\frac{n}{c_7k^2}}\end{aligned}$$

for constants c_i's > 1. Thus, the class-wise error $\mathbb{P}_{e|c}$ is upper bounded by

$$(k-1)\exp\left(-\|\boldsymbol{\mu}\|_2^2 \frac{\left(\left(1 - \frac{C_1}{\sqrt{n}} - \frac{C_2 n}{p}\right)\|\boldsymbol{\mu}\|_2 - C_3\min\{\sqrt{k}, \sqrt{\log(2n)}\}\right)^2}{C_4\left(\|\boldsymbol{\mu}\|_2^2 + \frac{kp}{n}\right)}\right)$$

with probability at least $1 - \frac{c_4}{n} - c_8 k e^{-\frac{n}{c_7k^2}}$. This completes the proof. □

C.1.5 Proof of Equation (68)

Here, using the results of Section F.1, we show how to obtain Equation (68) from Equation (65). First, by [WT21, Appendix C.2] (with $\mathbf{y}$ replaced by $\mathbf{v}_m$), we have

$$\mathbf{v}_m(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_m = \frac{s_{mm}^{(-m)}}{\det_{-m}}, \quad \text{for all} \quad m \in [k],$$

where $\det_{-m} = (\|\boldsymbol{\mu}_m\|_2^2 - t_{mm}^{(-m)})s_{mm}^{(-m)} + (h_{mm}^{(-m)} + 1)^2$. Then [WT21, Equation (44)] gives

$$\|\boldsymbol{\mu}_c\|_2^2 \cdot \mathbf{v}_c(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_c + \mathbf{v}_c(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{d}_c = \frac{\|\boldsymbol{\mu}_c\|_2^2 s_{cc}^{(j)} - s_{cc}^{(j)}t_{cc}^{(j)} + {h_{cc}^{(j)}}^2 + h_{cc}^{(j)}}{\det_j},$$

where $\det_j = (\|\boldsymbol{\mu}\|_2^2 - t_{cc}^{(j)})s_{cc}^{(j)} + (h_{cc}^{(j)}+1)^2$. Note that $\det_j = \det_{-c}$ when $c = k$ and $j = k-1$.

For $\mathbf{v}_c(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_j$ and $\mathbf{v}_j(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{d}_c$, we can again express the k-th order quadratic forms in terms of j-th order quadratic forms as follows:

$$\mathbf{v}_c(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_j = \frac{s_{cj}^{(j)} + s_{cj}^{(j)}h_{cc}^{(j)} - s_{cc}^{(j)}h_{jc}^{(j)}}{\det_j},$$
$$\mathbf{v}_j(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{d}_c = \frac{\|\boldsymbol{\mu}\|_2^2 s_{cc}^{(j)}h_{jc}^{(j)} - \|\boldsymbol{\mu}\|_2^2 s_{cj}^{(j)}h_{cc}^{(j)} + h_{cc}^{(j)}h_{jc}^{(j)} + h_{jc}^{(j)} - s_{cj}^{(j)}t_{cc}^{(j)}}{\det_j}.$$

Thus, we have

$$\|\boldsymbol{\mu}_c\|_2^2\mathbf{v}_c(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_j + \mathbf{v}_j(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{d}_c = \frac{\|\boldsymbol{\mu}_c\|_2^2 s_{jc}^{(j)} + h_{jc}^{(j)} + h_{jc}^{(j)}h_{cc}^{(j)} - s_{jc}^{(j)}t_{cc}^{(j)}}{\det_j}.$$

This completes the proof. □

C.2 Proof of Theorem 6

In this section we prove Theorem 6. The simplex ETF setting for the class means gives us $\|\boldsymbol{\mu}\|_2^2 = -(k-1)\boldsymbol{\mu}_m^T\boldsymbol{\mu}_c$ for $m \neq c$. Therefore, following the analysis above, the additional term $\sum_{m\neq c}\boldsymbol{\mu}_m^T\boldsymbol{\mu}_c\mathbf{v}_m^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_c$ is upper bounded by $\|\boldsymbol{\mu}\|_2^2 \max_{m,c}|\mathbf{v}_m^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_c|$. Since the dominating term in $\mathbf{v}_m^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_c$ is $s_{cj}^{(j)}$, which has a much smaller upper bound than the dominating term in $\mathbf{v}_c^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_c$, and the term $\boldsymbol{\mu}_j^T\boldsymbol{\mu}_c\mathbf{v}_j^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{v}_j$ in $\widehat{\mathbf{w}}_j^T\boldsymbol{\mu}_c$ has a positive contribution to $\widehat{\mathbf{w}}_c^T\boldsymbol{\mu}_c - \widehat{\mathbf{w}}_j^T\boldsymbol{\mu}_c$ under the simplex ETF setting, the final generalization bound does not change except up to constant factors.

C.3 Proof of Corollary 3

We now prove the condition for benign overfitting provided in Corollary 3. Note that following Theorem 2, we assume that

$$p > C_1k^3n\log(kn) + n - 1 \quad \text{and} \quad p > C_2k^{1.5}n^{1.5}\|\boldsymbol{\mu}\|_2. \tag{73}$$

We begin with the setting where $\|\boldsymbol{\mu}\|_2^2 > C\frac{kp}{n}$, for some $C > 1$. In this case, we get that Equation (72) is lower bounded by $\frac{1}{c}\left(\left(1 - \frac{C_3}{\sqrt{n}} - \frac{C_4n}{p}\right)\|\boldsymbol{\mu}\|_2 - C_5\sqrt{k}\right)^2$, and we have

$$\begin{aligned}\left(\left(1 - \frac{C_3}{\sqrt{n}} - \frac{C_4n}{p}\right)\|\boldsymbol{\mu}\|_2 - C_5\sqrt{k}\right)^2 &> \|\boldsymbol{\mu}\|_2^2 - 2\|\boldsymbol{\mu}\|_2^2\frac{C_3}{\sqrt{n}} - 2\|\boldsymbol{\mu}\|_2^2\frac{C_4n}{p} - 2C_5\sqrt{k}\|\boldsymbol{\mu}\|_2 \\ &> \left(1 - \frac{2C_3}{\sqrt{n}}\right)\frac{kp}{n} - 2\|\boldsymbol{\mu}\|_2^2\frac{C_4n}{p} - 2C_5\sqrt{k}\|\boldsymbol{\mu}\|_2. \end{aligned} \tag{74}$$

Then Equation (73) gives

$$\begin{aligned}(74) &> \left(1 - \frac{2C_3}{\sqrt{n}}\right)\frac{kp}{n} - \left(\frac{p}{k^{1.5}n^{1.5}}\right)^2\frac{C_6n}{p} - \frac{C_7\sqrt{k}p}{k^{1.5}n^{1.5}} \\ &= \frac{kp}{n}\left(1 - \frac{2C_3}{\sqrt{n}} - \frac{C_6}{k^4n} - \frac{C_7}{k^2\sqrt{n}}\right), \end{aligned} \tag{75}$$

which goes to $+\infty$ as $\left(\frac{p}{n}\right) \to \infty$.

Next, we consider the case $\|\boldsymbol{\mu}\|_2^2 \leq \frac{kp}{n}$. Moreover, we assume that $\|\boldsymbol{\mu}\|_2^4 = C_2\left(\frac{p}{n}\right)^\alpha$, for $\alpha > 1$. Then, Equation (72) is lower bounded by $\frac{n}{ckp}\|\boldsymbol{\mu}\|_2^4\left(\left(1 - \frac{C_3}{\sqrt{n}} - \frac{C_4n}{p}\right) - \frac{C_5\sqrt{k}}{\|\boldsymbol{\mu}\|_2}\right)^2$, and we get

$$\begin{aligned}\frac{n}{kp}\|\boldsymbol{\mu}\|_2^4\left(\left(1 - \frac{C_3}{\sqrt{n}} - \frac{C_4n}{p}\right) - \frac{C_5\sqrt{k}}{\|\boldsymbol{\mu}\|_2}\right)^2 &> \left(1 - \frac{2C_3}{\sqrt{n}}\right)\frac{n}{kp}\|\boldsymbol{\mu}\|_2^4 - \frac{C_6n^2}{kp^2}\|\boldsymbol{\mu}\|_2^4 - \frac{C_7n}{\sqrt{kp}}\|\boldsymbol{\mu}\|_2^3 \\ &\geq \left(1 - \frac{2C_3}{\sqrt{n}}\right)\frac{1}{k}\left(\frac{p}{n}\right)^{\alpha-1} - \frac{C_6}{k}\left(\frac{p}{n}\right)^{\alpha-2} - \frac{C_7}{\sqrt{k}}\left(\frac{p}{n}\right)^{0.75\alpha-1}, \end{aligned} \tag{76}$$

where the last inequality uses Equations (73) and condition $\|\boldsymbol{\mu}\|_2^2 \le \frac{kp}{n}$. Consequently, the RHS of Equation (76) will go to $+\infty$ as $\left(\frac{p}{n}\right) \to \infty$, provided that $\alpha > 1$. Overall, it suffices to have

$$p > \max\left\{C_1 k^3 n \log(kn) + n - 1, C_2 k^{1.5} n^{1.5} \|\boldsymbol{\mu}\|_2, \frac{n\|\boldsymbol{\mu}\|_2^2}{k}\right\},$$
$$\text{and } \|\boldsymbol{\mu}\|_2^4 \ge C_8 \left(\frac{p}{n}\right)^{\alpha}, \quad \text{for } \alpha \in (1, 2].$$

All of these inequalities hold provided that $\|\boldsymbol{\mu}\|_2 = \Theta(p^{\beta})$ for $\beta \in (1/4, 1/2]$ for finite k and n. This completes the proof. □

D Main lemmas used in error analysis of MLM

In this section, we collect the proofs of the main lemmas that are used in the error analysis of MLM (proof of Theorem 5, provided in Section 6.3). We first introduce notation that is specific to these proofs.

For two indices $\ell, j \in [k]$, we use the Kronecker delta notation $\delta_{\ell,j} = \mathbb{I}[\ell \neq j]$. For a diagonal covariance matrix $\boldsymbol{\Sigma}$ and $\ell \ge 1$, we define the leave-ℓ-out covariance matrix $\boldsymbol{\Sigma}_{-1:\ell}$ as $\boldsymbol{\Sigma}$ with the first ℓ rows and columns removed. For an arbitrary PSD matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ with eigenvalues $\lambda_1, \ldots, \lambda_d$ and any index $k \in \{0, \ldots, d-1\}$, the effective rank of the first kind (in the sense of [BLLT20]) is defined as

$$r_k(\boldsymbol{M}) := \frac{1}{\lambda_{k+1}} \cdot \sum_{\ell=k+1}^{d} \lambda_\ell. \tag{77}$$

Additionally, we state our convention for constants for this proof. Hereafter, we let $c, C \ldots > 0$ denote positive absolute constants in lower and upper bounds respectively. We also use $c_k, C_k > 0$ in a similar manner to denote constants that may only depend on the number of classes k. To simplify exposition in the proof, the values of these constants may be changing from line to line without explicit reference. Finally, by "large enough" n we mean that $n \ge C_k$ for some universal constant C_k that depends only on k.

D.1 Proof of Lemma 4

The proof of Lemma 4 follows similarly to the proof of Theorem 4 in Appendix D.3 and Lemma 11 in Appendix E of [MNS+21], with the two important and nontrivial extensions mentioned above: one, to the multiclass case involving k signal vectors, and two, considering the logistic model for label noise. Without loss of generality[8], we assume for simplicity that $j_c = c$ for all $c \in [k]$, and consider classes $c_1 = 1, c_2 = 2$ for the argument. First, we consider the following adjusted orthonormal basis

$$\widetilde{\boldsymbol{e}}_1 = \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2}, \qquad \widetilde{\boldsymbol{e}}_2 = \frac{(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2}, \qquad \widetilde{\boldsymbol{e}}_j = \boldsymbol{e}_j \text{ for all } j \ge 3.$$

This orthonormal basis together with the bilevel ensemble structure in Definition 4 then gives us

$$\begin{aligned}\mathsf{SU}_{1,2} &= \frac{\sqrt{\lambda_H}}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2} \cdot (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top (\widehat{\mathbf{w}}_1 - \widehat{\mathbf{w}}_2) = \sqrt{\lambda_H} \cdot \widetilde{\boldsymbol{e}}_1^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{v}_1 - \mathbf{v}_2) \\ &= \sqrt{\lambda_H} \cdot \widetilde{\boldsymbol{e}}_1^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{y}_1.\end{aligned}$$

where in the second line we introduce the shorthand $\mathbf{y}_1 := \mathbf{v}_1 - \mathbf{v}_2$. Next, we define

$$\mathbf{A} := \mathbf{X}^\top \mathbf{X} = \sum_{j=1}^{p} \lambda_j \mathbf{z}_j \mathbf{z}_j^\top,$$

[8]The reason this is without loss of generality is because we can carry out the same analysis otherwise with the appropriate permutation of the index labels.

with $\mathbf{z}_j := \frac{1}{\sqrt{\lambda_j}}\mathbf{X}^\top \widetilde{\boldsymbol{e}}_j, j = 1, \ldots, p$ and note that $\mathbf{z}_j \overset{\text{iid}}{\sim} \mathcal{N}(0, I_n)$. (This uses again the rotational invariance of Gaussianity and the bilevel ensemble structure.) Finally, for $\ell = 1, \ldots, p-1$, we denote the "leave-ℓ-out" matrices *corresponding to the changed basis* by $\mathbf{A}_{-1:\ell} := \sum_{j=\ell+1}^{p} \lambda_j \mathbf{z}_j \mathbf{z}_j^\top$. Note that, by definition, $\mathbf{A}_{-1:0} := \mathbf{A}$.

Using the above notation, we can then write the survival terms as follows

$$\mathsf{SU}_{1,2} = \lambda_H \cdot \mathbf{z}_1^\top \mathbf{A}^{-1} \mathbf{y}_1.$$

The main challenge in characterizing the term above is that $\mathbf{A}^{-1}$ is dependent on both $\mathbf{z}_1$ and $\mathbf{y}_1$. In particular, $\mathbf{y}_1$ depends on $\mathbf{z}_1$ itself but also it depends on $\mathbf{z}_2, \ldots, \mathbf{z}_k$. In the binary case, a simple *leave-one-out* analysis suffices to circumvent this difficulty as shown in [MNS+21]. In the multiclass setting, we need to do a much more challenging *leave-k-out* analysis which we outline below. In particular, we outline a recursive argument over k steps that iteratively removes the dependencies on $\mathbf{z}_1, \ldots, \mathbf{z}_k$ from $\mathbf{A}^{-1}$. This process is described in the following subsections.

D.1.1 Key recursion: Removing dependencies

Start by defining the following "quadratic-like" terms:

$$Q_\ell := \mathbf{y}_1^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_1, \qquad \ell = 0, \ldots, k\,, \tag{78a}$$

$$\widetilde{Q}_\ell := \mathbf{z}_1^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_1, \qquad \ell = 1, \ldots, k\,, \tag{78b}$$

$$R_{\ell,j} := \mathbf{y}_1^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_j, \qquad \ell \geq j, j = 1, \ldots, k-1\,. \tag{78c}$$

Recall the term we wish to control is $Q_0 = \mathbf{z}_1^T \mathbf{A}_{-1:0}^{-1} \mathbf{y}_1 = \mathbf{z}_1^T \mathbf{A}^{-1} \mathbf{y}_1$. A single application of the matrix inversion lemma (which was also done in [MNS+21] for the binary case and is described in a self-contained manner in Appendix D.1.2) yields that $\mathsf{SU}_{1,2} = \frac{\lambda_H Q_1}{1+\lambda_H \widetilde{Q}_1}$. However, unlike in the binary case Q_1 can no longer be easily controlled, as $\mathbf{y}_1$ still depends on $\mathbf{A}_{-1:1}^{-1}$ as it is a functional of not only $\mathbf{z}_1$, but also $\{\mathbf{z}_2, \ldots, \mathbf{z}_k\}$.

On the other hand, the term involving the leave-k-out Gram matrix, i.e. $Q_k = \mathbf{z}_1^T \mathbf{A}_{-1:k}^{-1} \mathbf{y}_1$ avoids this issue. This is because $\mathbf{y}_1$ is only a functional of $\{\mathbf{z}_1, \ldots, \mathbf{z}_k\}$, which ensures that $\mathbf{y}_1$ is independent of $\mathbf{A}_{-1:k}^{-1}$. This allows us to sharply characterize Q_k via the Hanson-Wright inequality, as shown in the lemma below.

Lemma 15. *For large enough n, we have*

$$c_k \left(\frac{cn}{\lambda_L r_s(\boldsymbol{\Sigma})} + \frac{c' n^{3/4}}{\lambda_L r_s(\boldsymbol{\Sigma})} \right) \geq Q_k \geq c_k \left(\frac{(n-s)}{c\lambda_L r_s(\boldsymbol{\Sigma})} - \frac{c' n^{3/4}}{\lambda_L r_s(\boldsymbol{\Sigma})} \right). \tag{79}$$

with probability at least $1 - 2e^{-\sqrt{n}}$.

See Appendix E.2.1 for the proof of this lemma.

Thus, it suffices to characterize Q_1 in terms of Q_k, so that we can translate upper/lower bounds on Q_k to upper/lower bounds on Q_1 and thereby characterize the survival $\mathsf{SU}_{1,2}$. The main result of this section, shown below, does precisely this, guaranteeing that $|Q_1 - Q_k| = o(Q_k)$ with high probability.

Lemma 16. *We have*

$$\left(1 - \frac{C_k}{n^{1/4}}\right) Q_k \leq Q_1 \leq \left(1 + \frac{C_k}{n^{1/4}}\right) Q_k.$$

with probability at least $1 - C' k^3 e^{-C\sqrt{n}}$.

In the remainder of this section we prove Lemma 16. We introduce the following recursion for any $\ell = k, \ldots, 1$ by directly applying the matrix inversion lemma:

$$\begin{aligned} Q_{\ell-1} = \mathbf{z}_1^T \mathbf{A}_{-1:\ell-1}^{-1} \mathbf{y}_1 = \mathbf{z}_1^T (\mathbf{A}_{-1:\ell} + \mathbf{z}_\ell \mathbf{z}_\ell^T)^{-1} \mathbf{y}_1 = Q_\ell - \frac{\lambda_H (\mathbf{z}_1^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell)(\mathbf{y}_1^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell)}{1 + \lambda_H \mathbf{z}_\ell^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell} \\ = Q_\ell - \widetilde{Q}_\ell \left(\frac{\mathbf{z}_1^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell}{\widetilde{Q}_\ell} \right) \left(\frac{\lambda_H R_{\ell,\ell}}{1 + \lambda_H \mathbf{z}_\ell^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell} \right), \end{aligned} \tag{80}$$

where in the last line we recalled the definitions of $\widetilde{Q}_\ell$ and of $R_{\ell,\ell}$ in Eqs. (78).

In order to prove Lemma 16 using the above recursion, we establish the following bounds on each of the three terms $\frac{\mathbf{z}_1^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell}{\widetilde{Q}_\ell}$, $R_{\ell,\ell}$ and $\widetilde{Q}_\ell$ that appear in Eq. (80). We provide the proofs of each of these technical lemmas in Appendix E.2.

Lemma 17. *For large enough n and for all $\ell \in [k]$, we have*

$$|\mathbf{z}_1^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell| \le \frac{C}{n^{1/4}} \mathbf{z}_1^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_1 = \frac{C}{n^{1/4}} \widetilde{Q}_\ell \,, \tag{81}$$

with probability at least $1 - Ck^3 e^{-\sqrt{n}}$.

Lemma 18. *For all $\ell \in [k]$, we have*

$$|R_{\ell,\ell}| \le C_k \cdot \mathbf{z}_\ell^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell \,, \quad \ell = k, k-1, \dots, 1 \,. \tag{82}$$

with probability at least $1 - Ck^3 e^{-\sqrt{n}}$.

Lemma 19. *For large enough n and for all $\ell \in [k]$, we have*

$$0 \le \widetilde{Q}_\ell \le \frac{2}{c_k} Q_k \tag{83}$$

with probability at least $1 - Cke^{-\sqrt{n}}$.

Proof of Lemma 16 Combining the bounds in Lemmas 17, 18, 19 and the fact that $\mathbf{z}_\ell^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell \ge 0$ within Equation (80) immediately yields for all $\ell = 1, \dots, k$:

$$\begin{aligned} |Q_\ell - Q_{\ell-1}| &\le \frac{2}{c_k} Q_k \left(\frac{C}{n^{1/4}} \right) \left(\frac{C_k \cdot \lambda_H \mathbf{z}_\ell^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell}{1 + \lambda_H \mathbf{z}_\ell^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell} \right) \\ &\le Q_k \frac{C_k}{c_k n^{1/4}} . \end{aligned}$$

The desired then follows by the bound $|Q_k - Q_1| \le \sum_{\ell=2}^k |Q_\ell - Q_{\ell-1}|$.

D.1.2 Completing the proof of Lemma 4

Armed with Lemma 16, we now complete the proof of Lemma 4. Recall that

$$\mathsf{SU}_{1,2} = \lambda_H Q_0 = \lambda_H \cdot \mathbf{z}_1^\top \mathbf{A}^{-1} \mathbf{y}_1 \,.$$

Applying Equation (80) for $\ell = 1$, we can write

$$\mathsf{SU}_{1,2} = \lambda_H Q_0 = \lambda_H \left(Q_1 - \frac{\lambda_H \widetilde{Q}_1 Q_1}{1 + \lambda_H \lambda_H \widetilde{Q}_1} \right) = \frac{\lambda_H Q_1}{1 + \lambda_H \widetilde{Q}_1} \,.$$

Thus, combining Lemmas 19 and 16, we can obtain the following lower/upper bounds on $\mathsf{SU}_{1,2}$:

$$\frac{\lambda_H \left(1 - \frac{C_k}{n^{1/4}}\right) Q_k}{1 + \lambda_H \left(\frac{2}{c_k}\right) Q_k} \le \mathsf{SU}_{1,2} \le \lambda_H \left(1 + \frac{C_k}{n^{1/4}}\right) Q_k. \tag{84}$$

It remains to substitute the upper/lower bounds on Q_k we obtained in Lemma 15. Plugging in the definition of the bilevel ensemble gives $\lambda_L r_s(\mathbf{\Sigma}) = n^m - n^r$. Noting that $m > 1$ and $r < 1$ gives

$$\begin{aligned} &\frac{cn}{\lambda_L r_s(\mathbf{\Sigma})} + \frac{c' n^{3/4}}{\lambda_L r_s(\mathbf{\Sigma})} \le C n^{1-m} \text{ and} \\ &\frac{(n-s)}{c \lambda_L r_s(\mathbf{\Sigma})} - \frac{c' n^{3/4}}{\lambda_L r_s(\mathbf{\Sigma})} \ge c n^{1-m}. \end{aligned}$$

Therefore, we have

$$cn^{1-m} \le Q_k \le Cn^{1-m}.$$

Noting that $\lambda_H = n^{m-q-r}$, we then have $cn^{1-q-r} \le \lambda_H Q_k \le Cn^{1-q-r}$. Plugging this back into Equation (84) gives

$$c_k n^{1-q-r} \le \mathsf{SU}_{1,2} \le C_k n^{1-q-r}$$

for large enough n, which is the desired statement. A union bound over Lemmas 15 and 16 implies that this statement holds with probability at least $1 - Ck^3 e^{-C\sqrt{n}}$. This completes the proof of Lemma 4. □

D.2 Proof of Lemma 5

This proof extends the argument in [MNS+21, Proof of Theorem 24] using the same change-of-basis argument that we used to characterize the survival. As with the proof of Lemma 4, we assume without loss of generality that $c_1 = 1, c_2 = 2$. First, we recall that

$$\widehat{\boldsymbol{\Delta}}_{1,2} := \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{v}_1 - \mathbf{v}_2) = \mathbf{X}\mathbf{A}^{-1}\mathbf{y}_1, \tag{85}$$

and that we defined $\mathbf{A} := \mathbf{X}^T\mathbf{X}$ and $\mathbf{y}_1 := \mathbf{v}_1 - \mathbf{v}_2$ as shorthand. We first state and prove the following lemma, which is analogous to [MNS+21, Lemma 28, Eq. (53a)].

Lemma 20. *The contamination term* $\mathsf{CN}_{1,2}$ *can be expressed as,*

$$\mathsf{CN}_{1,2} = \sqrt{\mathbf{y}_1^\top \mathbf{C}\mathbf{y}_1}, \text{ where} \tag{86}$$
$$\mathbf{C} := \mathbf{A}^{-1}\left(\sum_{j=1,j\neq 1}^{d} \lambda_j^2 \mathbf{z}_j\mathbf{z}_j^\top\right)\mathbf{A}^{-1}.$$

This is a consequence of the relation $\mathsf{CN}_{1,2}^2 := \sum_{j=1,j\neq 1}^{d} \lambda_j \hat{\alpha}_j^2$, *where we define* $\hat{\alpha}_j := \sqrt{\lambda_j}\cdot \mathbf{z}_j^\top \mathbf{A}^{-1}\mathbf{y}_j$.

See Appendix E.3 for the proof of this lemma.

Note that the expression in Lemma 20 is still challenging to characterize, as the difference of label vectors $\mathbf{y}_1$ is dependent on the matrix $\mathbf{C}$. To make progress, we will write a k-step recursive equation to express $\hat{\alpha}_j$ (and, thereby, $\mathsf{CN}_{1,2}$) in terms of $\mathbf{A}_{-1:k}^{-1}$ instead of $\mathbf{A}^{-1}$, leading to a possible characterization in terms of quadratic forms for which we can apply the Hanson-Wright inequality. We begin by reproducing the first recursion from the proof of [MNS+21, Lemma 28], which directly yields

$$\hat{\alpha}_j = \sqrt{\lambda_j}\cdot \mathbf{z}_j^\top \mathbf{A}_{-1:1}^{-1}(\mathbf{y}_1 - \mathsf{SU}_{1,2}\mathbf{z}_1).$$

We now recurse this argument to get an expression in terms of $\mathbf{A}_{-1:2}^{-1}$. Applying the Sherman-Morrison formula yields

$$\mathbf{A}_{-1:1}^{-1} = \mathbf{A}_{-1:2}^{-1} - \frac{\lambda_H \cdot \mathbf{A}_{-1:2}^{-1}\mathbf{z}_2\mathbf{z}_2\mathbf{A}_{-1:2}^{-1}}{1 + \lambda_H \cdot \mathbf{z}_2^\top \mathbf{A}_{-1:2}^{-1}\mathbf{z}_2}$$

and, consequently,

$$\hat{\alpha}_j = \sqrt{\lambda_j}\cdot \mathbf{z}_j^\top \mathbf{A}_{-1:2}^{-1}\left(\tilde{\mathbf{y}}_1 - \mathbf{z}_2 \cdot \frac{\lambda_H \cdot \mathbf{z}_2^\top \mathbf{A}_{-1:2}^{-1}\tilde{\mathbf{y}}_1}{1 + \lambda_H \cdot \mathbf{z}_2^\top \mathbf{A}_{-1:2}^{-1}\mathbf{z}_2}\right). \tag{87}$$

To write the entire k-step recursion, we define some shorthand notation. For $\ell = 2, \ldots, k$ we define

$$\mathsf{SU}_{1,2}^{(\ell)} := \frac{\lambda_H \cdot \mathbf{z}_\ell^\top \mathbf{A}_{-1:\ell}^{-1}\tilde{\mathbf{y}}_{\ell-1}}{1 + \lambda_H \cdot \mathbf{z}_\ell^\top \mathbf{A}_{-1:\ell}^{-1}\mathbf{z}_\ell} \text{ and}$$
$$\tilde{\mathbf{y}}_\ell := \tilde{\mathbf{y}}_{\ell-1} - \mathsf{SU}_{1,2}^{(\ell)}\mathbf{z}_\ell$$
$$\implies \tilde{\mathbf{y}}_k = \mathbf{y}_1 - \sum_{\ell=1}^{k} \mathsf{SU}_{1,2}^{(\ell)}\mathbf{z}_\ell.$$

Consequently, rewriting Equation (87) in terms of this shorthand notation gives

$$\hat{\alpha}_j = \sqrt{\lambda_j} \cdot \mathbf{z}_j^\top \mathbf{A}_{-1:2}^{-1} \tilde{\mathbf{y}}_2,$$

and repeating this argument for $\ell = 3, \dots, k$ ultimately yields

$$\hat{\alpha}_j = \sqrt{\lambda_j} \cdot \mathbf{z}_j^\top \mathbf{A}_{-1:k}^{-1} \tilde{\mathbf{y}}_k.$$

Then, we use an identical set of manipulations to the proof of [MNS+21, Lemma 28] (reproduced for completeness) to get

$$\begin{aligned}
\mathsf{CN}_{1,2}^2 = \sum_{j=1, j\neq 1}^{d} \lambda_j \hat{\alpha}_j^2 &= \sum_{j=1, j\neq 1}^{d} \lambda_j^2 \tilde{\mathbf{y}}_k^\top \mathbf{A}_{-1:k}^{-1} \mathbf{z}_j \mathbf{z}_j^\top \mathbf{A}_{-1:k}^{-1} \tilde{\mathbf{y}}_k \\
&= \tilde{\mathbf{y}}_k^\top \mathbf{A}_{-1:k}^{-1} \left(\sum_{j=1, j\neq 1}^{d} \lambda_j^2 \mathbf{z}_j \mathbf{z}_j^\top \right) \mathbf{A}_{-1:k}^{-1} \tilde{\mathbf{y}}_k \\
&= \tilde{\mathbf{y}}_k^\top \widetilde{\mathbf{C}}_k \tilde{\mathbf{y}}_k, \text{ where} \\
\widetilde{\mathbf{C}}_k &:= \mathbf{A}_{-1:k}^{-1} \left(\sum_{j=1, j\neq 1}^{d} \lambda_j^2 \mathbf{z}_j \mathbf{z}_j^\top \right) \mathbf{A}_{-1:k}^{-1}.
\end{aligned}$$

We now complete the proof of Lemma 5 by working with the expression $\mathsf{CN}_{1,2}^2 = \tilde{\mathbf{y}}_k^\top \widetilde{\mathbf{C}}_k \tilde{\mathbf{y}}_k$. First, we note that we can write

$$\begin{aligned}
\mathsf{CN}_{1,2}^2 &= \tilde{\mathbf{y}}_k^\top \widetilde{\mathbf{C}}_{k,1} \tilde{\mathbf{y}}_k + \tilde{\mathbf{y}}_k^\top \widetilde{\mathbf{C}}_{k,2} \tilde{\mathbf{y}}_k \text{ where} \\
\widetilde{\mathbf{C}}_{k,1} &:= \mathbf{A}_{-1:k}^{-1} \left(\sum_{j=1, j\neq 1}^{k} \lambda_j^2 \mathbf{z}_j \mathbf{z}_j^\top \right) \mathbf{A}_{-1:k}^{-1} \text{ and} \\
\widetilde{\mathbf{C}}_{k,2} &:= \mathbf{A}_{-1:k}^{-1} \left(\sum_{j=k+1}^{d} \lambda_j^2 \mathbf{z}_j \mathbf{z}_j^\top \right) \mathbf{A}_{-1:k}^{-1}.
\end{aligned}$$

Then, we can sharply upper-bound the terms $T_1 := \tilde{\mathbf{y}}_k^\top \widetilde{\mathbf{C}}_{k,1} \tilde{\mathbf{y}}_k$ and $T_2 := \tilde{\mathbf{y}}_k^\top \widetilde{\mathbf{C}}_{k,2} \tilde{\mathbf{y}}_k$, which we do below beginning with the second term T_2.

We apply the algebraic identity $(\mathbf{x}-\mathbf{y})^\top \mathbf{M} (\mathbf{x}-\mathbf{y}) \leq 2(\mathbf{x}^\top \mathbf{M} \mathbf{x} + \mathbf{y}^\top \mathbf{M} \mathbf{y})$ $k-1$ times to get

$$\tilde{\mathbf{y}}_k^\top \widetilde{\mathbf{C}}_{k,2} \tilde{\mathbf{y}}_k \leq 2^{k-1} \left(\mathbf{y}_1^\top \widetilde{\mathbf{C}}_{k,2} \mathbf{y}_1 + \sum_{\ell=1}^{k} (\mathsf{SU}^{(\ell)})_{1,2}^2 \cdot \mathbf{z}_\ell^\top \widetilde{\mathbf{C}}_{k,2} \mathbf{z}_\ell \right).$$

We use the following technical lemma, which is proved in Appendix E.3.

Lemma 21. *For $\ell = 2, \dots, k$ we define*

$$\mathsf{SU}_{1,2}^{(\ell)} := \frac{\lambda_H \cdot \mathbf{z}_\ell^\top \mathbf{A}_{-1:\ell}^{-1} \tilde{\mathbf{y}}_{\ell-1}}{1 + \lambda_H \cdot \mathbf{z}_\ell^\top \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell},$$

where $\tilde{\mathbf{y}}_\ell := \tilde{\mathbf{y}}_{\ell-1} - \mathsf{SU}_{1,2}^{(\ell)} \mathbf{z}_\ell$ and $\mathsf{SU}_{1,2}^{(1)} := \mathsf{SU}_{1,2} = \frac{\lambda_H \cdot \mathbf{z}_1^\top \mathbf{A}_{-1}^{-1} \mathbf{y}_1}{1 + \lambda_H \cdot \mathbf{z}_1^\top \mathbf{A}_{-1}^{-1} \mathbf{z}_1}$. Then, for all $\ell = 2, \dots, k$ we have

$$|\mathsf{SU}_{1,2}^{(\ell)}| \leq \frac{C_k}{n^{1/4}} < C_k \,. \tag{88}$$

with probability at least $1 - Ck^2 e^{-\sqrt{n}}$.

Applying Lemma 21 thus gives

$$\tilde{\mathbf{y}}_k^\top \widetilde{\mathbf{C}}_{k,2} \tilde{\mathbf{y}}_k \le C_k \left(\mathbf{y}_1^\top \widetilde{\mathbf{C}}_{k,2} \mathbf{y}_1 + \sum_{\ell=1}^{k} \mathbf{z}_\ell^\top \widetilde{\mathbf{C}}_{k,2} \mathbf{z}_\ell \right).$$

Now, we note that the matrix $\widetilde{\mathbf{C}}_{k,2}$ only depends on $\{\mathbf{z}_j\}_{j=k+1}^{p}$ and is therefore independent of $\mathbf{y}_1$ as well as $\{\mathbf{z}_\ell\}_{\ell=1}^{k}$. Recall that each of $\{\mathbf{z}_\ell\}_{\ell=1}^{k}$ is isotropic Gaussian and that $\mathbf{y}_1$ is sub-Gaussian with uncorrelated components, i.e. $y_{1,i}^2 \le 1$ and $\mathbb{E}[y_{1,i} y_{c,i'}] = 0$ for $i \ne i' \in [n]$. Therefore, we can apply the Hanson-Wright inequality [RV+13] with the parameters stated in [MNS+21, Eq (44)] to get

$$\tilde{\mathbf{y}}_k^\top \widetilde{\mathbf{C}}_{k,2} \tilde{\mathbf{y}}_k \le C_k \cdot \mathrm{Tr}(\widetilde{\mathbf{C}}_{k,2}) \cdot \log n$$

with probability at least $\left(1 - \frac{1}{n}\right)$. We denote by $\{\tilde{\lambda}_j\}_{j=1}^{p-k}$ the diagonal entries of the leave-k-out covariance matrix $\mathbf{\Sigma}_{-1:k}$. A direct application of [MNS+21, Lemma 30] (which, in turn, is taken from [BLLT20, Lemma 11]) gives

$$\mathrm{Tr}(\widetilde{\mathbf{C}}_{k,2}) \le C \left(\frac{s-k}{n} + n \cdot \frac{\sum_{j=s-k+1}^{p-k} \tilde{\lambda}_j^2}{(\sum_{j>s-k+1}^{p-k} \tilde{\lambda}_j)^2} \right).$$

Then, substituting the bilevel ensemble parameterization in a manner identical to the proof of [MNS+21, Lemma 35] gives

$$T_2 \le C_k \cdot n^{-\min(m-1, 2q+r-1)} \cdot \log n. \tag{89}$$

for $q > 1 - r$.

Controlling the term $T_1 := \tilde{\mathbf{y}}_k^\top \widetilde{\mathbf{C}}_{k,1} \tilde{\mathbf{y}}_k$ Unfortunately, this term is more delicate than T_2, because the matrix $\widetilde{\mathbf{C}}_{k,1}$ intricately depends on $\mathbf{z}_2, \dots, \mathbf{z}_k$. However, we can unravel the expression back to get

$$\begin{aligned}
\tilde{\mathbf{y}}_k^\top \widetilde{\mathbf{C}}_{k,1} \tilde{\mathbf{y}}_k &= \sum_{j=2}^{k} \lambda_j^2 (\mathbf{z}_j^\top \mathbf{A}_{-1:k}^{-1} \tilde{\mathbf{y}}_k)^2 \\
&\le \sum_{j=2}^{k} \lambda_j^2 \left(|\mathbf{z}_j^\top \mathbf{A}_{-1:k}^{-1} \mathbf{y}_1| + \sum_{\ell=1}^{k} |\mathsf{SU}_{1,2}^{(\ell)}| |\mathbf{z}_j^\top \mathbf{A}_{-1:k}^{-1} \mathbf{z}_\ell| \right)^2 \\
&\le C_k \sum_{j=2}^{k} \lambda_j^2 \left(|\mathbf{z}_j^\top \mathbf{A}_{-1:k}^{-1} \mathbf{y}_1| + |\mathbf{z}_j^\top \mathbf{A}_{-1:k}^{-1} \mathbf{z}_1| + \frac{1}{n^{1/4}} \sum_{\ell=2}^{k} |\mathbf{z}_j^\top \mathbf{A}_{-1:k}^{-1} \mathbf{z}_\ell| \right)^2
\end{aligned}$$

where the last inequality uses Lemma 21 and Lemma 4.

The key observation is that there are only $O(k^2)$ such terms that we need to control. Noting that $\mathbf{A}_{-1:k}^{-1}$ is independent of each of $\mathbf{y}_1$ and $\{\mathbf{z}_j\}_{j=1}^{k}$, we now use the Hanson-Wright inequality to control each of the terms $\{\mathbf{z}_j^\top \mathbf{A}_{-1:k}^{-1} \mathbf{y}_1\}_{j=2}^{k}$ and $\{\mathbf{z}_j^\top \mathbf{A}_{-1:k}^{-1} \mathbf{z}_\ell\}_{j \ne \ell}$. Note that for $j = 2, \dots, k$, we have $\mathbb{E}[\mathbf{y}_1 \mathbf{z}_j^\top] = \mathbf{0}$ from Lemma 22 (a base technical lemma, proved in Appendix E.1) and $\mathbb{E}[\mathbf{z}_\ell \mathbf{z}_j^\top] = \delta_{\ell,j} \mathbf{I}_p$. We apply this inequality (as stated in [MNS+21, Lemma 26]) for the choice $t = \|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot \sqrt{n \log n}$ to get

$$\begin{aligned}
&\mathbf{z}_j^\top \mathbf{A}_{-1:k}^{-1} \mathbf{y}_1 \le \|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot \sqrt{n \log n} \text{ and} \\
&\mathbf{z}_j^\top \mathbf{A}_{-1:k}^{-1} \mathbf{z}_\ell \le \delta_{\ell,j} \cdot \mathrm{tr}(\mathbf{A}_{-1:k}^{-1}) + \|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot \sqrt{n \log n},
\end{aligned}$$

each with probability at least $1 - \frac{1}{n^c}$ for some $c > 0$. Next, applying Lemma 23 (a base technical lemma, proved in Appendix E.1) gives us $\|\mathbf{A}_{-1:k}^{-1}\|_2 \le \frac{C}{\lambda_L r_s(\mathbf{\Sigma})}$ with probability at least $1 - 2e^{-\frac{n}{c}}$ over the random matrix $\mathbf{A}_{-1:k}^{-1}$. We further recall that $\lambda_L r_s(\mathbf{\Sigma}) = n^m - n^r \ge c n^m$ for large enough n, and

that $\lambda_j = \lambda_H = n^{m-q-r}$ for $j = 2, \ldots, k$ (because under our assumptions $s > k$). Excluding the terms $\{\mathbf{z}_j^\top \mathbf{A}_{-1:k}^{-1} \mathbf{z}_j\}_{j=2}^k$ for now, each of the above contributes the following to T_1:

$$\frac{C_k \cdot \lambda_H^2 \cdot n \log n}{\lambda_L^2 r_s^2(\mathbf{\Sigma}_{-1:k})} \leq C_k \cdot n^{1-2q-2r} \cdot \log n =: C_k \cdot n^{-(2q+2r-1)} \cdot \log n < n^{-(2q+r-1)},$$

which is identical to the scaling for T_2. We finally return to controlling the terms $\{\mathbf{z}_j^\top \mathbf{A}_{-1:k}^{-1} \mathbf{z}_j\}_{j=2}^k$. Note that each of these terms is pre-multiplied by the factor $\frac{1}{n^{1/4}}$ Applying Lemma 23 again gives $\mathrm{tr}(\mathbf{A}_{-1:k}^{-1}) \leq \frac{Cn}{\lambda_L r_s(\mathbf{\Sigma})}$ with probability at least $1 - 2e^{-\frac{n}{c}}$. The contribution from each of these terms, thus, becomes

$$\begin{aligned}\frac{1}{n^{1/2}} \frac{C_k \lambda_H^2 n^2}{\lambda_L^2 r_s^2(\mathbf{\Sigma}_{-1:k})} + \frac{1}{n^{1/2}} \frac{C_k \cdot \lambda_H^2 \cdot n \log n}{\lambda_L^2 r_s^2(\mathbf{\Sigma}_{-1:k})} &\leq \frac{1}{n^{1/2}} \frac{C_k \lambda_H^2 n^2}{\lambda_L^2 r_s^2(\mathbf{\Sigma}_{-1:k})} \\ &\leq C_k \cdot n^{2-2q-2r-1/2} = C_k \cdot n^{-(2q+2r-3/2)}.\end{aligned}$$

Thus, we get

$$T_1 \leq C_k \cdot n^{-\min(2q+r-1, 2q+2r-3/2)} \cdot \log n. \tag{90}$$

Putting it all together Recall that $\mathsf{CN}_{1,2}^2 := T_1 + T_2$. Therefore, putting together the upper bounds from Equations (90) and (89) gives us the following statement:

$$\mathsf{CN}_{1,2}(n) \leq C_k \sqrt{\log n} \cdot n^{-\frac{\min\{m-1, 2q+r-1, 2q+2r-3/2\}}{2}}$$

for $q > 1 - r$ and a universal constant C_k that depends only on k. This is the desired statement. Further, a union bound over each of the probabilistic inequalities implies that the statement holds with probability at least $1 - \frac{C_k}{n^c}$ for some $0 < c \leq 1$. This completes the proof of Lemma 5. □

E Supporting technical lemmas for MLM error analysis

In this section, we prove the supporting technical lemmas for the MLM error analysis.

E.1 Basic lemmas about the MLM

We begin by collecting basic lemmas about the MLM that form building blocks to prove the rest of the technical lemmas. The first such basic lemma controls the expectation of certain product forms involving the difference label vector $\mathbf{y}_1$ and individual feature vectors $\{\mathbf{z}_\ell\}_{\ell=1}^p$.

Lemma 22. *Let $\mathbf{y}_1 = \mathbf{v}_1 - \mathbf{v}_2$ be the difference label vector for $c_1 = 1, c_2 = 2$ and $\{\mathbf{z}_\ell\}_{\ell=1}^p$ be defined as in the proof of Lemma 4. Then, we have for every $i \in [n]$,*

$$c_{k,\ell} := \mathbb{E}[y_{1,i} z_{\ell,i}] = c_k \delta_{1,\ell},$$

where $c_k > 0$ is a universal positive constant that depends only on k.

Proof. To prove this lemma we utilize the orthogonality and equal-weight Assumption 5 as well as the details of the MLM. We denote $\mathbf{u}_j := \mathbf{X}^\top \boldsymbol{e}_j$. It is easy to see from the definition of the changed basis $\{\mathbf{z}_j\}_{j=1}^p$ that $\mathbf{z}_j = \mathbf{u}_j$ for all $j \geq 3$, and $\mathbf{z}_1 = \frac{1}{\sqrt{2}}(\mathbf{u}_1 - \mathbf{u}_2)$ and $\mathbf{z}_2 = \frac{1}{\sqrt{2}}(\mathbf{u}_1 + \mathbf{u}_2)$. We now use the simplex-ETF-type structure of $\mathbf{v}_1, \mathbf{v}_2$ together with the structure in the MLM model to get

$$\begin{aligned}\mathbb{P}\left(y_{1,i} = 1 \middle| \{u_{1,i}, u_{2,i}, \ldots, u_{k,i}\}\right) &= \frac{\exp(u_{1,i})}{\sum_{c' \in [k]} \exp(u_{c',i})} \text{ and} \\ \mathbb{P}\left(y_{1,i} = -1 \middle| \{u_{1,i}, u_{2,i}, \ldots, u_{k,i}\}\right) &= \frac{\exp(u_{2,i})}{\sum_{c' \in [k]} \exp(u_{j_{c',i}})},\end{aligned}$$

and $y_{1,i} = 0$ otherwise. Note here that $\{u_{c,i}\}_{c\in[k]}$ are i.i.d. standard Gaussian. We start with the case $\ell = 1$. Here, we get

$$\begin{aligned}
\mathbb{E}[z_{1,i}y_{1,i}] &= \frac{1}{\sqrt{2}} \cdot \mathbb{E}\left[(u_{1,i} - u_{2,i}) \cdot \frac{\exp(u_{1,i})}{\sum_{c'\in[k]} \exp(u_{c',i})} - (u_{1,i} - u_{2,i}) \cdot \frac{\exp(u_{2,i})}{\sum_{c'\in[k]} \exp(u_{j_{c'},i})}\right] \\
&= \frac{1}{\sqrt{2}} \cdot \mathbb{E}\left[(U_1 - U_2) \cdot \frac{(e^{U_1} - e^{U_2})}{\sum_{c=1}^k e^{U_c}}\right] \\
&= \sqrt{2} \cdot \mathbb{E}\left[U_1 \cdot \frac{(e^{U_1} - e^{U_2})}{\sum_{c=1}^k e^{U_c}}\right],
\end{aligned}$$

where the last step follows by symmetry. Note that we have overloaded notation and written $U_c := u_{c,i}$ for each $c \in [k]$. We also write $\mathbf{U} := \begin{bmatrix} U_1 & \dots & U_k \end{bmatrix}$ as shorthand. Because U_c i.i.d. $\sim \mathcal{N}(0,1)$, we have

$$c_k = \mathbb{E}\left[U_1 \cdot g(\mathbf{U})\right]$$

where $g(\mathbf{U}) := \frac{e^{U_1} - e^{U_2}}{\sum_{c=1}^k e^{U_c}}$. Then, applying Stein's lemma, we get

$$\begin{aligned}
\mathbb{E}\left[U_1 \cdot g(\mathbf{U})\right] &= \sum_{i=1}^n \mathbb{E}[U_1 U_i] \cdot \mathbb{E}\left[\frac{\partial g}{\partial U_i}\right] \\
&= \mathbb{E}\left[\frac{\partial g}{\partial U_1}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i\geq 3} e^{U_1+U_i} + 2e^{U_1+U_2}}{(\sum_{i=1}^k e^{U_i})^2}\right] =: c_k > 0.
\end{aligned}$$

The last step follows because the argument inside the expectation can never take value 0 and is always non-negative. Thus, we have proved that $\mathbb{E}[y_{1,i}z_{1,i}] = c_k > 0$.

We now prove that $\mathbb{E}[y_{1,i}z_{\ell,i}] = 0$ for $\ell \neq 1$. First, for $\ell \geq 3$, we have

$$\mathbb{E}[y_{1,i}z_{\ell,i}] = \frac{1}{\sqrt{2}} \cdot \mathbb{E}\left[u_{\ell,i} \cdot \frac{\exp(u_{1,i})}{\sum_{c'\in[k]} \exp(u_{c',i})} - u_{\ell,i} \cdot \frac{\exp(u_{2,i})}{\sum_{c'\in[k]} \exp(u_{c',i})}\right] = 0$$

by symmetry. Next, for $\ell = 2$, we have

$$\begin{aligned}
\mathbb{E}[z_{2,i}y_{1,i}] &= \frac{1}{\sqrt{2}} \cdot \mathbb{E}\left[(u_{1,i} + u_{2,i}) \cdot \frac{\exp(u_{1,i})}{\sum_{c'\in[k]} \exp(u_{c',i})} - (u_{1,i} + u_{2,i}) \cdot \frac{\exp(u_{2,i})}{\sum_{c'\in[k]} \exp(u_{j_{c'},i})}\right] \\
&= \frac{1}{\sqrt{2}} \cdot \mathbb{E}\left[(U_1 + U_2) \cdot \frac{(e^{U_1} - e^{U_2})}{\sum_{c=1}^k e^{U_c}}\right] \\
&= \mathbb{E}\left[U_1 \cdot \frac{(e^{U_1} - e^{U_2})}{\sum_{c=1}^k e^{U_c}}\right] - \mathbb{E}\left[U_2 \cdot \frac{(e^{U_2} - e^{U_1})}{\sum_{c=1}^k e^{U_c}}\right] = 0,
\end{aligned}$$

where the last equality follows by symmetry. This completes the proof. □

The next basic lemma controls the trace and operator norm of leave-ℓ-out Gram matrices and leverages ideas first appearing in [BLLT20].

Lemma 23. *For all $\ell \in [k]$ and sufficiently large n, the following inequalities are true for universal constants $c, C > 0$, each with probability at least $1 - 2e^{-\frac{n}{c}}$:*

$$\|\mathbf{A}_{-1:\ell}^{-1}\|_2 \leq \frac{c}{\lambda_L r_s(\mathbf{\Sigma})}$$

and

$$\frac{cn}{\lambda_L r_s(\mathbf{\Sigma})} \geq \mathsf{tr}(\mathbf{A}_{-1:\ell}^{-1}) \geq \frac{(n-s)}{c\lambda_L r_s(\mathbf{\Sigma})}.$$

In particular, these imply

$$\frac{\|\mathbf{A}_{-1:\ell}^{-1}\|_2 \cdot n^{3/4}}{\mathsf{tr}(\mathbf{A}_{1:\ell}^{-1})} \le \frac{C_2}{n^{1/4}}\,. \tag{91}$$

Proof. First, we upper bound the operator norm term. Observe that

$$\|\mathbf{A}_{-1:\ell}^{-1}\|_2 = \mu_1(\mathbf{A}_{-1:\ell}^{-1}) = \frac{1}{\mu_n(\mathbf{A}_{-1:\ell})} \le \frac{1}{\mu_n(\mathbf{A}_{-1:s})} \le \frac{c}{\lambda_{s+1} r_s(\mathbf{\Sigma})},$$

where the last inequality uses [BLLT20, Lemma 5]. The second-to-last inequality holds for any choice of $s > k \ge \ell$.

Next, we prove the bounds for the trace term. We lower bound the trace term as

$$\mathsf{tr}(\mathbf{A}_{-1:\ell}^{-1}) = \sum_{j=1}^{n} \frac{1}{\mu_j(\mathbf{A}_{-1:\ell})} \ge \sum_{j=s}^{n} \frac{1}{\mu_j(\mathbf{A}_{-1:\ell})} \ge \frac{(n-s)}{\mu_{s+1}(\mathbf{A}_{-1:\ell})}.$$

Thus, it remains to upper bound $\mu_{s+1}(\mathbf{A}_{-1:\ell})$. Let $\{\widetilde{\lambda}_j\}_{j=1}^{p-\ell}$ denote the re-indexed eigenvalues of $\mathbf{\Sigma}_{-1:\ell}$. Then, Equation (38) from Lemma 25 in [MNS+21] directly yields

$$\mu_{s+1}(\mathbf{A}_{-1:\ell}) \le C\widetilde{\lambda}_{s+1} r_s(\mathbf{\Sigma}_{-1:\ell})$$

provided that $r_s(\mathbf{\Sigma}_{-1:\ell}) \ge bn$. (Note that, under the bilevel ensemble, we have $r_s(\mathbf{\Sigma}_{-1:\ell}) = \frac{n^m - \ell - s}{\lambda_L} \ge cn^m > bn$ for large enough n.) Similarly, we upper bound the trace term as

$$\mathsf{tr}(\mathbf{A}_{-1:\ell}^{-1}) \le \frac{n}{\mu_n(\mathbf{A}_{-1:\ell})} \le \frac{cn}{\widetilde{\lambda}_{s+1} r_s(\mathbf{\Sigma}_{-1:\ell})}$$

where we now used Equation (37) from Lemma 25 in [MNS+21]. To complete the proof for the trace term, we show that $\widetilde{\lambda}_{s+1} r_s(\mathbf{\Sigma}_{-1:\ell}) \asymp \lambda_L r_s(\mathbf{\Sigma})$. First, we note that $\widetilde{\lambda}_{s+1} = \lambda_{s+1} = \lambda_L$ under the bilevel ensemble. Also recall that $\ell \le k < s$; hence we have $r_s(\mathbf{\Sigma}_{-1:\ell}) = p - s - \ell$ and $r_s(\mathbf{\Sigma}) = p - s$, which implies that $r_s(\mathbf{\Sigma}_{-1:\ell}) \asymp r_s(\mathbf{\Sigma})$ for large enough n. Putting all of this together yields the desired inequalities about the trace.

Finally, we prove Equation (91). This follows because, as already shown, we have

$$\|\mathbf{A}_{-1:\ell}^{-1}\|_2 \cdot n^{3/4} \le \frac{c\, n^{3/4}}{\lambda_L r_s(\mathbf{\Sigma})} \mathsf{tr}(\mathbf{A}_{-1:\ell}^{-1}) \ge \frac{n-s}{c\,\lambda_L r_s(\mathbf{\Sigma})}.$$

thereby giving us

$$\frac{\|\mathbf{A}_{-1:\ell}^{-1}\|_2 \cdot n^{3/4}}{\mathsf{tr}(\mathbf{A}_{-1:\ell}^{-1})} \le \frac{n^{3/4}}{n-s} \le \frac{2}{n^{1/4}}\,, \tag{92}$$

where the last inequality follows for large enough n because $s = n^r$ and we have assumed $r < 1$. This completes the proof of the lemma. □

The following basic lemma relates the ratios of quadratic forms that are "similar" in their probability distribution.

Lemma 24. *We have*

$$\frac{\mathbf{z}_\ell^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_\ell}{\mathbf{z}_{\ell'}^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_{\ell'}} \le C \qquad \textit{and} \qquad \frac{\mathbf{z}_\ell^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell}{\mathbf{z}_{\ell'}^T \mathbf{A}_{-1:\ell'}^{-1} \mathbf{z}_{\ell'}} \le C$$

for all $\ell, \ell' \in [k]$ with probability at least $1 - c\,k\,e^{-\sqrt{n}}$.

Proof. Recall that for any ℓ, ℓ', we have that $\mathbf{z}_\ell, \mathbf{z}_{\ell'}$ are both independent of $\mathbf{A}_{-1:k}^{-1}$. Therefore, we have

$$\mathbf{z}_{\ell'}^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_{\ell'} \geq \mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) - c_1 \|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4} \text{ and}$$
$$\mathbf{z}_\ell^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_\ell \leq \mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) + c_1 \|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}$$

with probability at least $1 - ke^{-\sqrt{n}}$. Putting these together gives

$$\frac{\mathbf{z}_\ell^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_\ell}{\mathbf{z}_{\ell'}^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_{\ell'}} \leq \frac{\mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) + c_1 \|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}}{\mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) - c_1 \|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}} \leq \frac{1 + \frac{c_1}{n^{1/4}}}{1 - \frac{c_1}{n^{1/4}}} \leq 2\,,$$

for large enough n, where in the above we used Eq. (91).

To prove the second inequality, recall that for any $\ell, \ell' \in [k]$, $\mathbf{z}_\ell$ is independent of $\mathbf{A}_{-1:\ell}^{-1}$ and $\mathbf{z}_{\ell'}$ is independent of $\mathbf{A}_{-1:\ell'}^{-1}$. Consequently, we have

$$\mathbf{z}_{\ell'}^T \mathbf{A}_{-1:\ell'}^{-1} \mathbf{z}_{\ell'} \geq \mathsf{tr}(\mathbf{A}_{-1:\ell'}^{-1}) - c_1 \|\mathbf{A}_{-1:\ell'}^{-1}\|_2 \cdot n^{3/4} \text{ and}$$
$$\mathbf{z}_\ell^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell \leq \mathsf{tr}(\mathbf{A}_{-1:\ell}^{-1}) + c_1 \|\mathbf{A}_{-1:\ell}^{-1}\|_2 \cdot n^{3/4}$$

with probability at least $1 - ke^{-\sqrt{n}}$. Putting these together gives

$$\frac{\mathbf{z}_\ell^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell}{\mathbf{z}_{\ell'}^T \mathbf{A}_{-1:\ell'}^{-1} \mathbf{z}_{\ell'}} \leq \frac{\mathsf{tr}(\mathbf{A}_{-1:\ell}^{-1}) + c_1 \|\mathbf{A}_{-1:\ell}^{-1}\|_2 \cdot n^{3/4}}{\mathsf{tr}(\mathbf{A}_{-1:\ell'}^{-1}) - c_1 \|\mathbf{A}_{-1:\ell'}^{-1}\|_2 \cdot n^{3/4}} \leq \frac{1 + \frac{c_1}{n^{1/4}}}{1 - \frac{c_1}{n^{1/4}}} \leq 2\,,$$

for large enough n, where we again used Eq. (91). This completes the proof of the lemma. □

Finally, the following basic lemma controls the ratio of traces of the leave-ℓ-out Gram matrix and the leave-k-out Gram matrix for any $\ell \in [k]$.

Lemma 25. *For all $\ell \in [k]$ and sufficiently large n, it holds for universal constant C that*

$$\frac{\mathsf{tr}(\mathbf{A}_{-1:\ell}^{-1})}{\mathsf{tr}(\mathbf{A}_{1:k}^{-1})} \geq \left(1 - \frac{C}{n}\right)^{k-\ell} \geq \left(1 - \frac{C}{n}\right)^k$$

Proof. Fix any $\ell \in [k]$. We first lower-bound the ratio $\frac{\mathsf{tr}(\mathbf{A}_{-1:\ell)}^{-1})}{\mathsf{tr}(\mathbf{A}_{-1:\ell+1}^{-1})}$, and then apply the argument recursively. Since $\mathbf{A}_{-1:\ell)} = \mathbf{A}_{-1:\ell+1} + \lambda_H \mathbf{z}_{\ell+1} \mathbf{z}_{\ell+1}^T$, we can apply the matrix inversion lemma to get

$$\mathbf{A}_{-1:\ell}^{-1} = \mathbf{A}_{-1:\ell+1}^{-1} - \frac{\lambda_H \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1} \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1}}{1 + \lambda_H \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1}}\,.$$

Hence, we have

$$\begin{aligned}
\mathsf{tr}(\mathbf{A}_{-1:\ell}^{-1}) &= \mathsf{tr}(\mathbf{A}_{-1:\ell+1}^{-1}) - \frac{\lambda_H \,\mathsf{tr}(\mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1} \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1})}{1 + \lambda_H \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1}} = \mathsf{tr}(\mathbf{A}_{-1:\ell+1}^{-1}) - \frac{\lambda_H \,\mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-2} \mathbf{z}_{\ell+1}}{1 + \lambda_H \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1}} \\
&\geq \mathsf{tr}(\mathbf{A}_{-1:\ell+1}^{-1}) - \|\mathbf{A}_{-1:\ell+1}^{-1}\|_2 \cdot \frac{\lambda_H \,\mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1}}{1 + \lambda_H \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1}} \\
&\geq \mathsf{tr}(\mathbf{A}_{-1:\ell+1}^{-1}) - \|\mathbf{A}_{-1:\ell+1}^{-1}\|_2
\end{aligned}$$

(The second inequality follows because for any positive semidefinite matrix $\mathbf{M}$ with eigendecomposition $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i^T$ we have

$$\mathbf{x}^T \mathbf{M}^2 \mathbf{x} = (\mathbf{U}\mathbf{x})^T \mathbf{\Lambda}^2 (\mathbf{U}\mathbf{x}) = \sum_i \lambda_i^2 (\mathbf{u}_i^T \mathbf{x})^2 \leq \left(\max_i \lambda_i\right) \sum_i \lambda_i (\mathbf{u}_i^T \mathbf{x})^2 = \|\mathbf{M}\|_2 \cdot \mathbf{x}^T \mathbf{M} \mathbf{x}$$

for any vector $\mathbf{x}$.) Continuing from the penultimate display, we obtain

$$\frac{\mathsf{tr}(\mathbf{A}_{-1:\ell}^{-1})}{\mathsf{tr}(\mathbf{A}_{-1:\ell+1}^{-1})} \geq 1 - \frac{\|\mathbf{A}_{-1:\ell+1}^{-1}\|_2}{\mathsf{tr}(\mathbf{A}_{-1:\ell+1}^{-1})} \geq 1 - \frac{C}{n}$$

where the last inequality applies Eq. (91). Recursively applying the above for $\ell + 1, \ldots, k$ completes the proof of the lemma. □

E.2 Survival Term

In this section we provide the proofs of Lemmas 15, 17, 18 and 19.

E.2.1 Proof of Lemma 15

First, we note that $\mathbf{y}_1$ remains independent of $\mathbf{A}_{-1:k}$ as $\mathbf{y}_1$ only depends on $\mathbf{z}_1, \dots, \mathbf{z}_k$ (which are in turn mutually independent of $\mathbf{z}_{k+1}, \dots, \mathbf{z}_p$ which comprise of $\mathbf{A}_{-1:k}$). Therefore, we can directly apply the Hanson-Wright inequality to get

$$Q_k \geq c_k \cdot \sqrt{\frac{2}{\pi}}\mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) - 2c_1\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4} \text{ and} \tag{93a}$$

$$Q_k \leq c_k \cdot \sqrt{\frac{2}{\pi}}\mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) + 2c_1\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}. \tag{93b}$$

with probability at least $1 - 2e^{-\sqrt{n}}$. Combining the above with Lemma 23 applied for $\ell = k$ directly gives the desired statement of Equation (79), completing the proof of the lemma. □

E.2.2 Proof of Lemma 17

Note that the quadratic-like terms in both the LHS and RHS of (81) are well-suited for an application of the Hanson-Wright inequality, since $\mathbf{z}_1, \mathbf{z}_\ell$ are independent of $\mathbf{A}_{-1:\ell}^{-1}$ for all $\ell = 2, \dots, k$. This is formalized in the lemma below. Specifically, the desired statement to prove Lemma 17, i.e. Eq. (81) for $\ell = 1, \dots, k$, follows directly by applying Lemma 26 below for the special case $\ell' = j = 1$. (The slightly more general statement of the lemma below will prove useful for proving subsequent lemmas.)

Lemma 26. *For large enough n, for all $\ell \in [k]$ and $\ell' < \ell, j \leq \ell$ we have*

$$|\mathbf{z}_\ell^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_{\ell'}| \leq \frac{C}{n^{1/4}} \mathbf{z}_j^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_j \,.$$

with probability at least $1 - Ck^3 e^{-\sqrt{n}}$.

Proof. The key observation is that for all $\ell' < \ell, j \leq \ell$, we have that $\mathbf{z}_\ell, \mathbf{z}_{\ell'}$, and $\mathbf{z}_j$ are all mutually independent of $\mathbf{A}_{-1:\ell}^{-1}$. Therefore, applying the Hanson-Wright inequality in the form stated by [MNS+21] gives us the following: for all $\ell \in [k], \ell' < \ell, j \leq \ell$, we have

$$\begin{aligned} |\mathbf{z}_\ell^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_{\ell'}| &\leq 2c_1\|\mathbf{A}_{-1:\ell}^{-1}\|_2 \cdot n^{3/4} \qquad \text{and} \\ \mathbf{z}_j^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_j &\geq \mathsf{tr}(\mathbf{A}_{-1:\ell}^{-1}) - c_1\|\mathbf{A}_{-1:\ell}^{-1}\|_2 \cdot n^{3/4} \,, \end{aligned}$$

with probability at least $1 - Ck^3 e^{-\sqrt{n}}$. Above, we used the fact that $\mathbf{z}_\ell, \mathbf{z}_{\ell'}$ are independent. Therefore, to prove the desired it suffices to show that

$$\mathsf{tr}(\mathbf{A}_{1:\ell}^{-1}) \geq \frac{n^{1/4}}{C_2}\|\mathbf{A}_{-1:\ell}^{-1}\|_2 \cdot n^{3/4} \,. \tag{94}$$

This follows immediately from Eq. (91) in Lemma 23. This completes the proof. □

E.2.3 Proof of Lemma 18

Recall that $R_{\ell,\ell} := \mathbf{y}_1^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell$. Bounding this term is difficult because $\mathbf{y}_1$ depends on $\mathbf{A}_{-1:\ell}^{-1}$ for any $\ell < k$. The only "easy" case is for $\ell = k$ for which $\mathbf{y}_1$ is independent of $\mathbf{A}_{-1:k}^{-1}$. As a starting point, we exploit this independence to control the terms $R_{k,\ell} = \mathbf{y}_1^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_\ell$ for all $\ell \in [k]$, in the lemma below.

Lemma 27. *We have, for large enough n,*

$$|R_{k,\ell}| = |\mathbf{y}_1^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_\ell| \leq \frac{C_k}{n^{1/4}} \mathbf{z}_k^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_k \quad \textit{for any } \ell = 2, \dots, k \,,$$

and

$$|R_{k,1}| = |\mathbf{y}_1^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_1| \leq C_k \mathbf{z}_k^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_k$$

with probability at least $1 - Cke^{-\sqrt{n}}$.

Proof. Recall that all of $\mathbf{y}_1, \mathbf{z}_\ell, \mathbf{z}_k$ are independent of $\mathbf{A}_{-1:k}^{-1}$. Therefore, we can apply the Hanson-Wright inequality to the RHS of the above, as well as $R_{k,\ell}$ (using the parallelogram law in the latter case) to get

$$\begin{aligned}
\mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) - c_1\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4} \le \mathbf{z}_k^\top \mathbf{A}_{-1:k}^{-1}\mathbf{z}_k \le \mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) + c_1\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4} \\
c_{k,\ell} \cdot \mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) - 2c_1\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4} \le R_{k,\ell} \le c_{k,\ell} \cdot \mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) + 2c_1\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4},
\end{aligned}$$

with probability at least $1 - Cke^{-\sqrt{n}}$. Above, we define $c_{k,\ell} := \mathbb{E}[y_{1,i}z_{\ell,i}]$ (identically for any $i \in [n]$). There are then two cases:

1. $\ell = 1$: In this case we get $c_{k,\ell} =: c_k > 0$ from Lemma 22. Plugging this above gives

$$\begin{aligned}
\frac{R_{k,1}}{\mathbf{z}_k^T\mathbf{A}_{-1:k}^{-1}\mathbf{z}_k} &\le \frac{c_k \cdot \mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) + 2c_1\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}}{\mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) - c_1\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}} \\
&\le c_k \frac{1 + \frac{c_1}{n^{1/4}}}{1 - \frac{c_1}{n^{1/4}}} \le 2c_k =: C_k,
\end{aligned}$$

   where the second inequality follows from Eq. (91) in Lemma 23 and the last inequality follows for large enough n. Similarly, we have

$$\begin{aligned}
\frac{R_{k,1}}{\mathbf{z}_k^T\mathbf{A}_{-1:k}^{-1}\mathbf{z}_k} &\ge -\frac{2c_1\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}}{\mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) + c_1\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}} \\
&= -\frac{2c_1}{\frac{\mathsf{tr}(\mathbf{A}_{-1:k}^{-1})}{\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}} + c_1} \\
&\ge -\frac{2c_1}{\frac{n^{1/4}}{2} + c_1} \ge -\frac{C}{n^{1/4}},
\end{aligned}$$

   where the second-to-last inequality in the above again used Equation (91).

2. $\ell \neq 1$: In this case we have $c_{k,\ell} = 0$, again from Lemma 22. Plugging this above gives

$$\begin{aligned}
\frac{R_{k,\ell}}{\mathbf{z}_k^T\mathbf{A}_{-1:k}^{-1}\mathbf{z}_k} &\le \frac{2c_1\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}}{\mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) - c_1\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}} \\
&\le \frac{2c_1}{\frac{\mathsf{tr}(\mathbf{A}_{-1:k}^{-1})}{\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}} + c_1} \le \frac{2c_1}{\frac{n^{1/4}}{2} + c_1} \le \frac{C}{n^{1/4}},
\end{aligned}$$

   where the last inequality follows for large enough n. Similarly, we have

$$\begin{aligned}
\frac{R_{k,\ell}}{\mathbf{z}_k^T\mathbf{A}_{-1:k}^{-1}\mathbf{z}_k} &\ge -\frac{2c_1\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}}{\mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) + c_1\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}} \\
&= -\frac{2c_1}{\frac{\mathsf{tr}(\mathbf{A}_{-1:k}^{-1})}{\|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}} + c_1} \\
&\ge -\frac{2c_1}{\frac{n^{1/4}}{2} + c_1} \ge -\frac{C}{n^{1/4}}
\end{aligned}$$

   where in the penultimate line we again used Eq. (91).

□

We now build on the "base case" Lemma 27 to control the terms $R_{\ell,\ell}$ in a similar manner to $R_{k,\ell}$. In particular, we note that the desired Eq. (82) to prove Lemma 18 follows by applying the slightly more general lemma below for the case $\ell' = \ell$.

Lemma 28. *For all $\ell \in [k]$ and all $\ell' \le \ell$, we have*

$$|R_{\ell,\ell'}| \le \begin{cases} C_k \cdot \mathbf{z}_\ell^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell & \text{if } \ell' = 1 \\ \frac{C_k}{n^{1/4}} \cdot \mathbf{z}_\ell^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell & \text{if } \ell' \neq 1 \end{cases} \tag{95}$$

with probability at least $1 - ck^3 e^{-\sqrt{n}}$.

We complete the proof of Lemma 18 by proving Lemma 28, which we do in the next section.

E.2.4 Proof of Lemma 28

We will use recursion starting from $\ell = k$ to prove the desired statement for all $\ell = k-1, k-2, \ldots, 1$. Throughout, we condition on the events of Lemmas 24, 26, and 27. The key to allow proving the statement recursively is the following relation that follows by the matrix-inversion-lemma and holds for all $\ell' \le \ell$:

$$\begin{aligned} R_{\ell,\ell'} &= \mathbf{y}_1^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_{\ell'} = \mathbf{y}_1^T \left(\mathbf{A}_{-1:\ell+1} + \mathbf{z}_{\ell+1} \mathbf{z}_{\ell+1}^T \right)^{-1} \mathbf{z}_{\ell'} \\ &= \mathbf{y}_1^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell'} - \frac{\lambda_H \left(\mathbf{y}_1^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell'} \right) \left(\mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell'} \right)}{1 + \lambda_H \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1}} \\ &= R_{\ell+1,\ell'} - R_{\ell+1,\ell+1} \frac{\lambda_H \left(\mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell'} \right)}{1 + \lambda_H \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1}} . \end{aligned} \tag{96}$$

First we prove the statement for the base case $\ell = k-1$. For any $\ell' \le k-1$, Equation (96) gives us

$$R_{k-1,\ell'} = R_{k,\ell'} - R_{k,k} \frac{\lambda_H \left(\mathbf{z}_k^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_{\ell'} \right)}{1 + \lambda_H \mathbf{z}_k^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_k} .$$

Note that because $\ell' \le k-1$, we have $\ell' < k$. Thus, we can apply Lemma 26 to get

$$|\epsilon_{k,\ell'}| := \frac{\lambda_H |\left(\mathbf{z}_k^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_{\ell'} \right)|}{1 + \lambda_H \mathbf{z}_k^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_k} \le \frac{C}{n^{1/4}} \frac{\lambda_H \left(\mathbf{z}_k^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_k \right)}{1 + \lambda_H \mathbf{z}_k^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_k} \le \frac{C}{n^{1/4}} .$$

Also, by Lemma 27, we have

$$\begin{aligned} |R_{k,1}| &\le C_k \mathbf{z}_k^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_k \text{ and} \\ |R_{k,j}| &\le \frac{C_k}{n^{1/4}} \cdot \mathbf{z}_k^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_k \text{ for all } j = 2, \ldots, k. \end{aligned}$$

Combining the three displays above with the recursion in Equation (96) yields the following for large enough n:

$$\begin{aligned} |R_{k-1,1}| &\le C_k \left(1 + C n^{-1/4} \right) \cdot \mathbf{z}_k^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_k \le C_k \cdot \mathbf{z}_k^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_k \text{ and} \\ |R_{k-1,\ell'}| &\le C_k n^{-1/4} \left(1 + C \right) \cdot \mathbf{z}_k^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_k \le \frac{C_k}{n^{1/4}} \cdot \mathbf{z}_k^T \mathbf{A}_{-1:k}^{-1} \mathbf{z}_k \text{ for all } \ell' \in \{2, \ldots, k-1\}. \end{aligned}$$

Lemma 24 (applied for the pair $(k, k-1)$) then gives us the desired Equation (95) for $\ell = k-1$, i.e. $|R_{k-1,1}| \le C_k \cdot \mathbf{z}_{k-1}^T \mathbf{A}_{-1:k-1}^{-1} \mathbf{z}_{k-1}$ and $|R_{k-1,\ell'}| \le C_k n^{-1/4} \cdot \mathbf{z}_{k-1}^T \mathbf{A}_{-1:k-1}^{-1} \mathbf{z}_{k-1}$ for $\ell' = 2, \ldots, k-1$. The base case is therefore proved.

Next, we prove the inductive step. In particular, we assume that Equation (95) is true for $\ell + 1$ and use it to prove the claim for ℓ. Our starting point is, again, the recursive relation in Equation (96). Noting that $\ell' < \ell + 1$, we can again apply Lemma 26 to get

$$|\epsilon_{\ell+1,\ell'}| := \frac{\lambda_H |\left(\mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell'} \right)|}{1 + \lambda_H \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1}} \le \frac{C}{n^{1/4}} \frac{\lambda_H \left(\mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1} \right)}{1 + \lambda_H \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1}} \le \frac{C}{n^{1/4}} .$$

Also, by the induction hypothesis, we have

$$|R_{\ell+1,1}| \le C_k\, \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1} \text{ and}$$
$$|R_{\ell+1,j}| \le \frac{C_k}{n^{1/4}}\, \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1} \text{ for all } j = 2, \dots, k.$$

Note that the sharper second inequality above applies to the term $R_{\ell+1,\ell+1}$ because we always have $\ell + 1 \ge 2$. Combining the two displays above with the recursion in Equation (96) yields the following for large enough n:

$$|R_{\ell,1}| \le C_k \left(1 + Cn^{-1/4}\right) \cdot \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1} \le C_k \cdot \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1} \le C_k \cdot \mathbf{z}_{\ell}^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_{\ell}$$

Similarly, we have for all $\ell' = 2, \dots, \ell$,

$$|R_{\ell,\ell'}| \le C_k n^{-1/4} (1 + C) \cdot \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1} \le C_k n^{-1/4} \cdot \mathbf{z}_{\ell+1}^T \mathbf{A}_{-1:\ell+1}^{-1} \mathbf{z}_{\ell+1} \le C_k n^{-1/4} \cdot \mathbf{z}_{\ell}^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_{\ell}$$

In both cases above, the last inequality follows from Lemma 24. This completes the proof of Lemma 28, and therefore the proof of Lemma 18. □

E.2.5 Proof of Lemma 19

Recall the definitions

$$Q_k := \mathbf{z}_1^T \mathbf{A}_{-1:k}^{-1} \mathbf{y}_1 \quad \text{and} \quad \widetilde{Q}_\ell := \mathbf{z}_1^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_1.$$

Since $\widetilde{Q}_\ell$ is a quadratic form, we have $\widetilde{Q}_\ell \ge 0$ and so it suffices to upper bound $\widetilde{Q}_\ell$. Because $\mathbf{z}_1$ is independent of $\mathbf{A}_{-1:\ell}^{-1}$ for any $\ell = 1, \dots, k$, we can directly apply the Hanson-Wright inequality to get

$$\widetilde{Q}_\ell = \mathbf{z}_1^\top \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_1 \le \mathsf{tr}(\mathbf{A}_{-1:\ell}^{-1}) + c_1 \|\mathbf{A}_{-1:\ell}^{-1}\|_2 \cdot n^{3/4}$$

with probability $1 - Ce^{-\sqrt{n}}$. Similarly, applying the Hanson-Wright inequality to the term Q_k (see Eq.(93)) we also have

$$Q_k \ge c_k \cdot \mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) - 2c_1 \|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}.$$

with the same probability. Putting these together, we get

$$\begin{aligned}
\frac{\widetilde{Q}_\ell}{Q_k} &\le \frac{\mathsf{tr}(\mathbf{A}_{1:\ell}^{-1}) + c_1 \|\mathbf{A}_{-1:\ell}^{-1}\|_2 \cdot n^{3/4}}{c_k \cdot \mathsf{tr}(\mathbf{A}_{-1:k}^{-1}) - 2c_1 \|\mathbf{A}_{-1:k}^{-1}\|_2 \cdot n^{3/4}} \\
&\le \frac{\mathsf{tr}(\mathbf{A}_{-1:k}^{-1})}{\mathsf{tr}(\mathbf{A}_{1:\ell}^{-1})} \frac{1 + \frac{c_1}{n^{1/4}}}{c_k - \frac{2c_1}{n^{1/4}}} \\
&\le \left(1 - \frac{C}{n}\right)^{-k} \left(\frac{1 + \frac{c_1}{n^{1/4}}}{c_k - \frac{2c_1}{n^{1/4}}}\right) \\
&\le \frac{2}{c_k},
\end{aligned}$$

where the second inequality follows from Lemma 23 for ℓ and k (for large enough n) and the second-to-last inequality uses Lemma 25. The last inequality follows again assuming large enough n. This completes the proof of the lemma. □

E.3 Contamination Term

In this section we prove Lemmas 20 and 21.

E.3.1 Proof of Lemma 20

First, we note that the desired Equation (86) is a direct consequence of the expression $\mathsf{CN}_{1,2}^2 := \sum_{j=1,j\neq 1}^d \lambda_j \hat{\alpha}_j^2$, where

$$\hat{\alpha}_j := \sqrt{\lambda_j} \cdot \mathbf{z}_j^\top \mathbf{A}^{-1} \mathbf{y}_j.$$

Therefore, it suffices to show that $\mathsf{CN}_{c_1,c_2}^2 = \sum_{j=1,j\neq 1}^d \lambda_j \hat{\alpha}_j^2$.

We denote the error vector $\boldsymbol{\xi} := \widehat{\boldsymbol{\Delta}}_{1,2} - \frac{\widehat{\boldsymbol{\Delta}}_{1,2}^\top \boldsymbol{\Sigma} \boldsymbol{\Delta}_{1,2}}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Delta}_{1,2}\|_2^2} \boldsymbol{\Delta}_{1,2}$ as shorthand, and recall from Lemma 3 that we have $\mathsf{CN}_{1,2}^2 = \boldsymbol{\xi}^\top \boldsymbol{\Sigma} \boldsymbol{\xi}$. Further, we define $\widetilde{\boldsymbol{E}} := \begin{bmatrix} \widetilde{\boldsymbol{e}}_1 \\ \widetilde{\boldsymbol{e}}_2 \\ \vdots \\ \widetilde{\boldsymbol{e}}_d \end{bmatrix} \in \mathbb{R}^{d\times d}$ as the changed basis in matrix form, and we define $\widetilde{\boldsymbol{\xi}} := \widetilde{\boldsymbol{E}} \boldsymbol{\xi}$. Then, the desired follows from these two statements:

1. We have $\mathsf{CN}_{1,2}^2 = \widetilde{\boldsymbol{\xi}}^\top \boldsymbol{\Sigma} \widetilde{\boldsymbol{\xi}}$.
2. We have $\tilde{\xi}_1 = 0$ and $\tilde{\xi}_j = \hat{\alpha}_j$ for all $j = 2, \ldots, d$.

We complete the proof by proving statements 1 and 2 for the specific form of $\boldsymbol{\Sigma}$ admitted by the bilevel ensemble.

Proof of statement 1 We prove this statement for a generic vector $\mathbf{y} \in \mathbb{R}^d$. Consider the vector $\tilde{\mathbf{y}} := \widetilde{\boldsymbol{E}} \mathbf{y}$. We will show that $\tilde{\mathbf{y}}^\top \boldsymbol{\Sigma} \tilde{\mathbf{y}} = \mathbf{y}^\top \boldsymbol{\Sigma} \mathbf{y}$. Because $\boldsymbol{\Sigma}$ is a diagonal matrix, we have $\mathbf{y}^\top \boldsymbol{\Sigma} \mathbf{y} = \sum_{j=1}^d \lambda_j y_j^2$. Further, it is straightforward to show from the specific form of the changed basis $\widetilde{\boldsymbol{E}}$ that $\tilde{\mathbf{y}}_1 = \frac{y_1 - y_2}{\sqrt{2}}$, $\tilde{\mathbf{y}}_2 = \frac{y_1 + y_2}{\sqrt{2}}$, and $\tilde{\mathbf{y}}_j = y_j$ for $j = 3, \ldots, d$. Therefore, we have

$$\begin{aligned}
\lambda_1 \tilde{y}_1^2 + \lambda_2 \tilde{y}_2^2 &= \lambda_H (\tilde{y}_1^2 + \tilde{y}_2^2) \\
&= \lambda_H \left(\frac{y_1^2 + y_2^2 - 2y_1 y_2 + y_1^2 + y_2^2 + 2y_1 y_2}{2} \right) \\
&= \lambda_H (y_1^2 + y_2^2) = \lambda_1 y_1^2 + \lambda_2 y_2^2.
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
\tilde{\mathbf{y}}^\top \boldsymbol{\Sigma} \tilde{\mathbf{y}} &= \sum_{j=1}^d \lambda_j \tilde{y}_j^2 = \lambda_1 \tilde{y}_1^2 + \lambda_2 \tilde{y}_2^2 + \sum_{j=3}^d \lambda_j \tilde{y}_j^2 \\
&= \lambda_1 y_1^2 + \lambda_2 y_2^2 + \sum_{j=3}^d \lambda_j y_j^2 = \sum_{j=1}^d \lambda_j y_j^2 = \mathbf{y}^\top \boldsymbol{\Sigma} \mathbf{y}.
\end{aligned}$$

Hence, we have proved statement 1.

Proof of statement 2 First, note that $\widetilde{\boldsymbol{\xi}} = \widetilde{\boldsymbol{E}} \widehat{\boldsymbol{\Delta}}_{1,2} - \frac{\widehat{\boldsymbol{\Delta}}_{1,2}^\top \boldsymbol{\Sigma} \boldsymbol{\Delta}_{1,2}}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Delta}_{1,2}\|_2^2} \cdot \widetilde{\boldsymbol{E}} \boldsymbol{\Delta}_{1,2}$. Recall that $\boldsymbol{\Delta}_{1,2} \propto \widetilde{\boldsymbol{e}}_1$, and so, $\widetilde{\boldsymbol{E}} \boldsymbol{\Delta}_{1,2} \propto \boldsymbol{e}_1$. Next, simple algebra shows that

$$\begin{aligned}
(\widetilde{\boldsymbol{E}} \widehat{\boldsymbol{\Delta}}_{1,2})_j &= \boldsymbol{e}_j^\top \widetilde{\boldsymbol{E}} \widehat{\boldsymbol{\Delta}}_{1,2} \\
&= \widetilde{\boldsymbol{e}}_j^\top \widehat{\boldsymbol{\Delta}}_{1,2} = \widetilde{\boldsymbol{e}}_j^\top \mathbf{X} \mathbf{A}^{-1} \mathbf{y}_1 \\
&= \sqrt{\lambda_j} \mathbf{z}_j^\top \mathbf{A}^{-1} \mathbf{y}_1 =: \hat{\alpha}_j.
\end{aligned}$$

where the third equality recalls the definition of $\widehat{\boldsymbol{\Delta}}_{1,2}$ from Equation (85) and the second-to-last equality recalls the definition $\mathbf{z}_j := \frac{1}{\sqrt{\lambda_j}} \mathbf{X}^\top \widetilde{\boldsymbol{e}}_j$. Noting that, by definition, $(\widetilde{\boldsymbol{E}} \boldsymbol{\Delta}_{1,2})_j = 0$ for all $j \neq 1$, we have

thus shown that $\tilde{\xi}_j = \hat{\alpha}_j$ for all $j = 2, \ldots, d$. To complete the proof of statement 2, we need to show that $\tilde{\xi}_1 = 0$. Denote $\boldsymbol{\Delta}_{1,2} = \alpha \widetilde{\boldsymbol{e}}_1$ for some $\alpha > 0$ (as a consequence, we also have $\widetilde{\boldsymbol{E}}\boldsymbol{\Delta}_{1,2} = \alpha \boldsymbol{e}_1$). Then, it is equivalent to show that

$$\frac{\widehat{\boldsymbol{\Delta}}_{1,2}^\top \boldsymbol{\Sigma} \boldsymbol{\Delta}_{1,2}}{\|\boldsymbol{\Sigma}^{1/2} \boldsymbol{\Delta}_{1,2}\|_2^2} \cdot \alpha = \widetilde{\boldsymbol{e}}_1^\top \widehat{\boldsymbol{\Delta}}_{1,2}.$$

(Recall that $\widetilde{\boldsymbol{E}}\boldsymbol{\Delta}_{1,2} \propto \boldsymbol{e}_1$, so this equality suffices to show the desired.) Starting with the LHS of the above, we get

$$\begin{aligned} \boldsymbol{\Sigma}\boldsymbol{\Delta}_{1,2} &= \lambda_1 \alpha \widetilde{\boldsymbol{e}}_1, \text{ and} \\ \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Delta}_{1,2} &= \lambda_1^{1/2} \alpha \widetilde{\boldsymbol{e}}_1. \end{aligned}$$

Therefore, we have

$$\begin{aligned} \frac{\boldsymbol{\Sigma}\boldsymbol{\Delta}_{1,2}}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Delta}_{1,2}\|_2^2} &= \frac{1}{\alpha}\widetilde{\boldsymbol{e}}_1, \text{ and} \\ \frac{\widehat{\boldsymbol{\Delta}}_{1,2}^\top \boldsymbol{\Sigma}\boldsymbol{\Delta}_{1,2}}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Delta}_{1,2}\|_2^2} \cdot \alpha &= \widehat{\boldsymbol{\Delta}}_{1,2}^\top \widetilde{\boldsymbol{e}}_1. \end{aligned}$$

This completes the proof of statement 2.

With statements 1 and 2 proved, the proof of this lemma is complete. □

E.3.2 Proof of Lemma 21

We prove the lemma using induction on $\ell = 1, \ldots, k$. For the base case $\ell = 1$, we have shown in Lemma 4 that

$$|\mathsf{SU}_{1,2}^{(1)}| := \frac{\lambda_H \cdot |\mathbf{z}_1^\top \mathbf{A}_{-1:1}^{-1} \mathbf{y}_1|}{1 + \lambda_H \cdot \mathbf{z}_1^\top \mathbf{A}_{-1:1}^{-1} \mathbf{z}_1} = |\mathsf{SU}_{1,2}| \le C_k \,. \tag{97}$$

Now, we prove the inductive step. We fix $\ell > 1$ and assume, along with the base case (Equation (97)), that the statement is also true for $2, \ldots, \ell - 1$, i.e.

$$\forall j = 2, \ldots, \ell - 1, \qquad |\mathsf{SU}_{1,2}^{(j)}| := \frac{\lambda_H \cdot |\mathbf{z}_j^\top \mathbf{A}_{-1:j}^{-1} \tilde{\mathbf{y}}_{j-1}|}{1 + \lambda_H \cdot \mathbf{z}_j^\top \mathbf{A}_{-1:j}^{-1} \mathbf{z}_j} \le \frac{C_k}{n^{1/4}} < C_k. \tag{98}$$

(In fact, as we will see, we will only need to apply the weaker inequality $|\mathsf{SU}_{1,2}^{(j)}| \le C_k$.) We use Equation (98) to prove the desired statement for ℓ. Consider first the numerator in the definition of $\mathsf{SU}_{1,2}^{(\ell)}$, i.e. the term

$$\mathbf{z}_\ell^\top \mathbf{A}_{-1:\ell}^{-1} \tilde{\mathbf{y}}_{\ell-1} = \mathbf{z}_\ell^\top \mathbf{A}_{-1:\ell}^{-1} \left(\mathbf{y}_1 - \sum_{j=1}^{\ell-1} \mathsf{SU}_{c_1,c_2}^{(j)} \mathbf{z}_j \right) = \mathbf{z}_\ell^\top \mathbf{A}_{-1:\ell}^{-1} \mathbf{y}_1 - \sum_{j=1}^{\ell-1} \mathsf{SU}_{c_1,c_2}^{(j)} \cdot \mathbf{z}_\ell^\top \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_j$$

Recall that $\mathbf{z}_\ell^\top \mathbf{A}_{-1:\ell}^{-1} \mathbf{y}_1 = R_{\ell,\ell}$. Note that Lemma 18 shows for $\ell \ge 2$ that

$$|R_{\ell,\ell}| \le \frac{C_k}{n^{1/4}} \mathbf{z}_\ell \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell \,,$$

Also recall from Lemma 26 that for all $j < \ell$, we have

$$|\mathbf{z}_\ell^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_j| \le C n^{-1/4} \cdot \mathbf{z}_j^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_j \le C n^{-1/4} \cdot \mathbf{z}_\ell^T \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell \,,$$

where again, the second inequality uses Lemma 24.

Putting the above together, applying the triangle inequality and using the induction hypothesis (i.e. $|\mathsf{SU}_{1,2}^{(j)}| \leq C_k$) we conclude that

$$\begin{aligned}|\mathbf{z}_\ell^\top \mathbf{A}_{-1:\ell}^{-1} \tilde{\mathbf{y}}_{\ell-1}| &\leq C \cdot \mathbf{z}_\ell \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell \left(C_k \cdot n^{-1/4} + \ell \cdot C_k \cdot n^{-1/4} \right) \\ &\leq \frac{C_k}{n^{1/4}} \cdot \mathbf{z}_\ell \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell \,. \end{aligned} \tag{99}$$

This gives us

$$|\mathsf{SU}_{1,2}^{(\ell)}| := \frac{\lambda_H \cdot |\mathbf{z}_\ell^\top \mathbf{A}_{-1:\ell}^{-1} \tilde{\mathbf{y}}_{\ell-1}|}{1 + \lambda_H \cdot \mathbf{z}_\ell^\top \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell} \leq \frac{1}{n^{1/4}} \cdot \frac{\lambda_H \cdot C_k \cdot \mathbf{z}_\ell \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell}{1 + \lambda_H \cdot \mathbf{z}_\ell^\top \mathbf{A}_{-1:\ell}^{-1} \mathbf{z}_\ell} \leq \frac{C_k}{n^{1/4}} \tag{100}$$

for all $\ell \geq 2$. This completes the proof of the lemma.

F Recursive formulas for higher-order quadratic forms

We first show how quadratic forms involving the j-th order Gram matrix $\mathbf{A}_j^{-1}$ can be expressed using quadratic forms involving the $(j-1)$-th order Gram matrix $\mathbf{A}_{j-1}^{-1}$. For concreteness, we consider $j = 1$; identical expressions hold for any $j > 1$ with the only change being in the superscripts. Recall from Section 6.2 that we can write

$$\mathbf{A}_1 = \mathbf{A}_0 + \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 \mathbf{v}_1 & \mathbf{Q}^T \boldsymbol{\mu}_1 & \mathbf{v}_1 \end{bmatrix} \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 \mathbf{v}_1^T \\ \mathbf{v}_1^T \\ \boldsymbol{\mu}_1^T \mathbf{Q} \end{bmatrix} = \mathbf{Q}^T \mathbf{Q} + \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 \mathbf{v}_1 & \mathbf{d}_1 & \mathbf{v}_1 \end{bmatrix} \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 \mathbf{v}_1^T \\ \mathbf{v}_1^T \\ \mathbf{d}_1^T \end{bmatrix}.$$

The first step is to derive an expression for $\mathbf{A}_1^{-1}$. By the Woodbury identity [HJ12], we get

$$\mathbf{A}_1^{-1} = \mathbf{A}_0^{-1} - \mathbf{A}_0^{-1} \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 \mathbf{v}_1 & \mathbf{d}_1 & \mathbf{v}_1 \end{bmatrix} \left[\mathbf{I} + \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 \mathbf{v}_1^T \\ \mathbf{v}_1^T \\ \mathbf{d}_1^T \end{bmatrix} \mathbf{A}_0^{-1} \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 \mathbf{v}_1 & \mathbf{d}_1 & \mathbf{v}_1 \end{bmatrix} \right]^{-1} \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 \mathbf{v}_1^T \\ \mathbf{v}_1^T \\ \mathbf{d}_1^T \end{bmatrix} \mathbf{A}_0^{-1}. \tag{101}$$

We first compute the inverse of the 3×3 matrix $\mathbf{B} := \left[\mathbf{I} + \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 \mathbf{v}_1^T \\ \mathbf{v}_1^T \\ \mathbf{d}_1^T \end{bmatrix} \mathbf{A}_0^{-1} \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 \mathbf{v}_1 & \mathbf{d}_1 & \mathbf{v}_1 \end{bmatrix} \right]$.

Recalling our definitions of the terms $s_{mj}^{(c)}, h_{mj}^{(c)}$ and $t_{mj}^{(c)}$ in Equation (40) in Section 6.2, we have:

$$\mathbf{B} = \begin{bmatrix} 1 + \|\boldsymbol{\mu}_1\|_2^2 s_{11}^{(0)} & \|\boldsymbol{\mu}_1\|_2 h_{11}^{(0)} & \|\boldsymbol{\mu}_1\|_2 s_{11}^{(0)} \\ \|\boldsymbol{\mu}_1\|_2 s_{11}^{(0)} & 1 + h_{11}^{(0)} & s_{11}^{(0)} \\ \|\boldsymbol{\mu}_1\|_2 h_{11}^{(0)} & t_{11}^{(0)} & 1 + h_{11}^{(0)} \end{bmatrix}.$$

Recalling $\mathbf{B}^{-1} = \frac{1}{\det_0} \mathrm{adj}(\mathbf{B})$, where $\det_0$ is the determinant of $\mathbf{B}$ and $\mathrm{adj}(\mathbf{B})$ is the adjoint of $\mathbf{B}$, simple algebra gives us

$$\det_0 = s_{11}^{(0)} (\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}) + (h_{11}^{(0)} + 1)^2,$$

and

$$\mathrm{adj}(\mathbf{B}) = \begin{bmatrix} (h_{11}^{(0)} + 1)^2 - s_{11}^{(0)} t_{11}^{(0)} & \|\boldsymbol{\mu}_1\|_2 (s_{11}^{(0)} t_{11}^{(0)} - h_{11}^{(0)} - {h_{11}^{(0)}}^2) & -\|\boldsymbol{\mu}_1\|_2 s_{11}^{(0)} \\ -\|\boldsymbol{\mu}_1\|_2 s_{11}^{(0)} & h_{11}^{(0)} + 1 + \|\boldsymbol{\mu}_1\|_2^2 s_{11}^{(0)} & -s_{11}^{(0)} \\ \|\boldsymbol{\mu}_1\|_2 (s_{11}^{(0)} t_{11}^{(0)} - h_{11}^{(0)} - {h_{11}^{(0)}}^2) & \|\boldsymbol{\mu}_1\|_2^2 {h_{11}^{(0)}}^2 - t_{11}^{(0)} (1 + \|\boldsymbol{\mu}_1\|_2^2 s_{11}^{(0)}) & h_{11}^{(0)} + 1 + \|\boldsymbol{\mu}_1\|_2^2 s_{11}^{(0)} \end{bmatrix}.$$

We will now use these expressions to derive expressions for the 1-order quadratic forms that are used in Appendix A.2.

F.1 Expressions for 1-st order quadratic forms

We now show how quadratic forms of order 1 can be expressed as a function of quadratic forms of order 0. All of the expressions are derived as a consequence of plugging in the expression for $\mathbf{B}^{-1}$ together with elementary matrix algebra.

First, we have

$$s_{mk}^{(1)} = \mathbf{v}_m^T \mathbf{A}_1^{-1} \mathbf{v}_k = \mathbf{v}_m^T \mathbf{A}_0^{-1} \mathbf{v}_k - \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 s_{m1}^{(0)} & h_{m1}^{(0)} & s_{m1}^{(0)} \end{bmatrix} \frac{\text{adj}(\mathbf{B})}{\det_0} \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 s_{k1}^{(0)} \\ s_{k1}^{(0)} \\ h_{k1}^{(0)} \end{bmatrix}$$
$$= s_{mk}^{(0)} - \frac{1}{\det_0} (\star)_s^{(0)}, \tag{102}$$

where we define

$$(\star)_s^{(0)} := (\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}) s_{1k}^{(0)} s_{1m}^{(0)} + s_{1m}^{(0)} h_{k1}^{(0)} h_{11}^{(0)} + s_{1k}^{(0)} h_{m1}^{(0)} h_{11}^{(0)} - s_{11}^{(0)} h_{k1}^{(0)} h_{m1}^{(0)} + s_{1m}^{(0)} h_{k1}^{(0)} + s_{1k}^{(0)} h_{m1}^{(0)}.$$

Thus, for the case $m = k$ we have

$$s_{kk}^{(1)} = \mathbf{v}_k^T \mathbf{A}_1^{-1} \mathbf{v}_k = \mathbf{v}_k^T \mathbf{A}_0^{-1} \mathbf{v}_k - \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 s_{k1}^{(0)} & h_{k1}^{(0)} & s_{k1}^{(0)} \end{bmatrix} \frac{\text{adj}(\mathbf{B})}{\det_0} \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 s_{k1}^{(0)} \\ s_{k1}^{(0)} \\ h_{k1}^{(0)} \end{bmatrix}$$
$$= s_{kk}^{(0)} - \frac{1}{\det_0} \Big((\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}) {s_{1k}^{(0)}}^2 + 2 s_{1k}^{(0)} h_{k1}^{(0)} h_{11}^{(0)} - s_{11}^{(0)} {h_{k1}^{(0)}}^2 + 2 s_{1k}^{(0)} h_{k1}^{(0)} \Big). \tag{103}$$

Next, we have

$$h_{mk}^{(1)} = \mathbf{v}_m^T \mathbf{A}_1^{-1} \mathbf{d}_k = \mathbf{v}_m^T \mathbf{A}_0^{-1} \mathbf{d}_k - \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 s_{m1}^{(0)} & h_{m1}^{(0)} & s_{m1}^{(0)} \end{bmatrix} \frac{\text{adj}(\mathbf{B})}{\det_0} \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 h_{1k}^{(0)} \\ h_{1k}^{(0)} \\ t_{1k}^{(0)} \end{bmatrix}$$
$$= h_{mk}^{(0)} - \frac{1}{\det_0} (\star)_h^{(0)}, \tag{104}$$

where we define

$$(\star)_h^{(0)} = (\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}) s_{1m}^{(0)} h_{1k}^{(0)} + h_{m1}^{(0)} h_{1k}^{(0)} h_{11}^{(0)} + h_{m1}^{(0)} h_{1k}^{(0)} + s_{1m}^{(0)} t_{k1}^{(0)} + s_{1m}^{(0)} t_{k1}^{(0)} h_{11}^{(0)} - s_{11}^{(0)} t_{k1}^{(0)} h_{m1}^{(0)}.$$

Next, we have

$$t_{km}^{(1)} = \mathbf{d}_k^T \mathbf{A}_1^{-1} \mathbf{d}_m = \mathbf{d}_k^T \mathbf{A}_0^{-1} \mathbf{d}_m - \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 h_{1k}^{(0)} & t_{1k}^{(0)} & h_{1k}^{(0)} \end{bmatrix} \frac{\text{adj}(\mathbf{B})}{\det_0} \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 h_{1m}^{(0)} \\ h_{1m}^{(0)} \\ t_{1m}^{(0)} \end{bmatrix}$$
$$= t_{km}^{(0)} - \frac{1}{\det_0} (\star)_t^{(0)}, \tag{105}$$

where we define

$$(\star)_t^{(0)} = (\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}) h_{1m}^{(0)} h_{1k}^{(0)} + t_{m1}^{(0)} h_{1k}^{(0)} h_{11}^{(0)} + t_{k1}^{(0)} h_{1m}^{(0)} h_{11}^{(0)} + t_{1m}^{(0)} h_{1k}^{(0)} + t_{1k}^{(0)} h_{1m}^{(0)} - s_{11}^{(0)} t_{1m}^{(0)} t_{1k}^{(0)}.$$

Thus, for the case $m = k$ we have

$$t_{kk}^{(1)} = \mathbf{d}_k^T \mathbf{A}_1^{-1} \mathbf{d}_k = \mathbf{d}_k^T \mathbf{A}_0^{-1} \mathbf{d}_k - \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 h_{1k}^{(0)} & t_{1k}^{(0)} & h_{1k}^{(0)} \end{bmatrix} \frac{\text{adj}(\mathbf{B})}{\det_0} \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 h_{1k}^{(0)} \\ h_{1k}^{(0)} \\ t_{1k}^{(0)} \end{bmatrix}$$
$$= t_{kk}^{(0)} - \frac{1}{\det_0} \Big((\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}) {h_{1k}^{(0)}}^2 + 2 t_{1k}^{(0)} h_{1k}^{(0)} h_{11}^{(0)} - s_{11}^{(0)} {t_{1k}^{(0)}}^2 + 2 t_{1k}^{(0)} h_{1k}^{(0)} \Big). \tag{106}$$

Next, we have

$$f_{ki}^{(1)} = \mathbf{d}_k^T \mathbf{A}_1^{-1} \mathbf{e}_i = \mathbf{d}_k^T \mathbf{A}_0^{-1} \mathbf{e}_i - \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 h_{1k}^{(0)} & t_{1k}^{(0)} & h_{1k}^{(0)} \end{bmatrix} \frac{\operatorname{adj}(\mathbf{B})}{\det_0} \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 g_{1i}^{(0)} \\ g_{1i}^{(0)} \\ f_{1i}^{(0)} \end{bmatrix}$$

$$= f_{ki}^{(0)} - \frac{1}{\det_0} (\star)_f^{(0)}, \tag{107}$$

where we define

$$(\star)_f^{(0)} = (\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}) h_{1k}^{(0)} g_{1i}^{(0)} + t_{1k}^{(0)} g_{1i}^{(0)} + t_{1k}^{(0)} h_{11}^{(0)} g_{1i}^{(0)} + h_{1k}^{(0)} f_{1i}^{(0)} + h_{1k}^{(0)} h_{11}^{(0)} f_{1i}^{(0)} - s_{11}^{(0)} t_{1k}^{(0)} f_{1i}^{(0)}.$$

Finally, we have

$$g_{ji}^{(1)} = \mathbf{v}_j^T \mathbf{A}_1^{-1} \mathbf{e}_i = \mathbf{v}_j^T \mathbf{A}_0^{-1} \mathbf{e}_i - \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 s_{j1}^{(0)} & h_{j1}^{(0)} & s_{j1}^{(0)} \end{bmatrix} \frac{\operatorname{adj}(\mathbf{B})}{\det_0} \begin{bmatrix} \|\boldsymbol{\mu}_1\|_2 g_{1i}^{(0)} \\ g_{1i}^{(0)} \\ f_{1i}^{(0)} \end{bmatrix}$$

$$= g_{ji}^{(0)} - \frac{1}{\det_0} (\star)_{gj}^{(0)}, \tag{108}$$

where we define

$$(\star)_{gj}^{(0)} = (\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)}) s_{1j}^{(0)} g_{1i}^{(0)} + g_{1i}^{(0)} h_{11}^{(0)} h_{j1}^{(0)} + g_{1i}^{(0)} h_{j1}^{(0)} + s_{1j}^{(0)} f_{1i}^{(0)} + s_{1j}^{(0)} h_{11}^{(0)} f_{1i}^{(0)} - s_{11}^{(0)} h_{j1}^{(0)} f_{1i}^{(0)}.$$

□

G One-vs-all SVM

In this section, we derive conditions under which the OvA solutions $\mathbf{w}_{\mathrm{OvA},c}$ interpolate, i.e, all data points are support vectors in Equation (8).

G.1 Gaussian mixture model

As in the case of the multiclass SVM, we assume nearly equal priors and nearly equal energy on the class means (Assumption 1).

Theorem 7. *Assume that the training set follows a multiclass GMM with noise covariance $\boldsymbol{\Sigma} = \mathbf{I}_p$ and Assumption 1 holds. Then, there exist constants $c_1, c_2, c_3 > 1$ and $C_1, C_2 > 1$ such that the solutions of the OvA-SVM and MNI are identical with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$ provided that*

$$p > C_1 kn \log(kn) + n - 1 \quad \text{and} \quad p > C_2 n^{1.5} \|\boldsymbol{\mu}\|_2. \tag{109}$$

We can compare Equation (109) with the corresponding condition for multiclass SVM in Theorem 2 (Equation (16)). Observe that the right-hand-side of Equation (109) above does not scale with k, while the right-hand-side of Equation (16) scales with k as k^3. Otherwise, the scalings with n and energy of class means $\|\boldsymbol{\mu}\|_2$ are identical. This discrepancy with respect to k-dependence arises because the multiclass SVM is equivalent to the OvA-SVM in Equation (34) with unequal margins $1/k$ and $(k-1)/k$ (as we showed in Theorem 1).

Proof sketch. Recall from Section 6.2 that we derived conditions under which the multiclass SVM interpolates the training data by studying the related symmetric OvA-type classifier defined in Equation (15). Thus, this proof is similar to the proof of Theorem 2 provided in Section 6.2. The only difference is that the margins for the OvA-SVM are not $1/k$ and $(k-1)/k$, but 1 for all classes. Owing to the similarity between the arguments, we restrict ourselves to a proof sketch here.

Following Section 6.2 and Equation (46), we consider $y_i = k$. We will derive conditions under which the condition

$$\left((1+h_{kk}^{(-k)})g_{ki}^{(-k)} - s_{kk}^{(-k)} f_{ki}^{(-k)}\right) + C\sum_{j\neq k}\left((1+h_{jj}^{(-j)})g_{ji}^{(-j)} - s_{jj}^{(-j)} f_{ji}^{(-j)}\right) > 0, \tag{110}$$

holds with high probability for some $C > 1$. We define

$$\epsilon := \frac{n^{1.5}\|\boldsymbol{\mu}\|_2}{p} \leq \tau,$$

where τ is chosen to be a sufficiently small constant. Applying the same trick as in Lemma 2 (with the newly defined parameters ϵ and τ) gives us with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$:

$$\begin{aligned}
(110) &\geq \left(\left(1 - \frac{C_1\epsilon}{\sqrt{k}\sqrt{n}}\right)\left(1 - \frac{1}{C_2}\right)\frac{1}{p} - \frac{C_3\epsilon}{n}\cdot\frac{n}{kp}\right) - \frac{k}{C_4}\left(\left(1 + \frac{C_5\epsilon}{\sqrt{k}\sqrt{n}}\right)\frac{1}{kp} - \frac{C_6\epsilon}{n}\cdot\frac{n}{kp}\right) \\
&\geq \left(1 - \frac{1}{C_9} - \frac{C_{10}\epsilon}{\sqrt{k}\sqrt{n}} - \frac{C_{11}\epsilon}{k} - C_{12}\epsilon\right)\frac{1}{p} \\
&\geq \frac{1}{p}\left(1 - \frac{1}{C_9} - C_0\tau\right),
\end{aligned} \tag{111}$$

for some constants C_i's > 1. We used the fact that $|g_{ji}^{(0)}| \leq (1/C)(1/(kp))$ for $j \neq y_i$ with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$ provided that $p > C_1 kn\log(kn) + n - 1$, which is the first sufficient condition in the theorem statement. □

G.2 Multinomial logistic model

Recall that we defined the data covariance matrix $\boldsymbol{\Sigma} = \sum_{i=1}^p \lambda_i \mathbf{v}_i\mathbf{v}_i^T = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$ and its spectrum $\boldsymbol{\lambda} = \begin{bmatrix}\lambda_1 & \cdots & \lambda_p\end{bmatrix}$. We also defined the effective dimensions $d_2 := \frac{\|\boldsymbol{\lambda}\|_1^2}{\|\boldsymbol{\lambda}\|_2^2}$ and $d_\infty := \frac{\|\boldsymbol{\lambda}\|_1}{\|\boldsymbol{\lambda}\|_\infty}$.

The following result provides sufficient conditions under which the OvA SVM and MNI classifier have the same solution with high probability under the MLM.

Theorem 8. *Assume that the training set follows a multiclass MLM. There exist constants c and $C_1, C_2 > 1$ such that, if the following conditions hold:*

$$d_\infty > C_1 n\log(kn) \quad \textit{and} \quad d_2 > C_2(\log(kn) + n), \tag{112}$$

the solutions of the OvA-SVM and MNI are identical with probability at least $(1 - \frac{c}{n})$. In the special case of isotropic covariance, the same result holds provided that

$$p > 10n\log(\sqrt{k}n) + n - 1, \tag{113}$$

Comparing this result to the corresponding results in Theorems 3, we observe that k now only appears in the log function (as a result of k union bounds). Thus, the unequal $1/k$ and $(k-1)/k$ margins that appear in the multiclass-SVM make interpolation harder than with the OvA-SVM, just as in the GMM case.

Proof sketch. For the OvA SVM classifier, we need to solve k binary max-margin classification problems, hence the proof follows directly from [MNS+21, Theorem 1] and [HMX21, Theorem 1] by applying k union bounds. We omit the details for brevity. □

One-vs-one SVM

In this section, we first derive conditions under which the OvO solutions interpolate, i.e, all data points are support vectors. We then provide an upper bound on the classification error of the OvO solution.

In OvO classification, we solve $k(k-1)/2$ binary classification problems, e.g. for classes pair (c, j), we solve

$$\mathbf{w}_{\mathrm{OvO},(c,j)} := \arg\min_{\mathbf{w}} \|\mathbf{w}\|_2 \quad \text{sub. to} \quad \mathbf{w}^T\mathbf{x}_i \geq 1, \text{ if } \mathbf{y}_i = c; \quad \mathbf{w}^T\mathbf{x}_i \leq -1 \text{ if } \mathbf{y}_i = j, \ \forall i \in [n]. \tag{114}$$

Then we apply these $k(k-1)/2$ classifiers to a fresh sample and the class that got the highest $+1$ voting gets predicted.

We now present conditions under which every data point becomes a support vector over these $k(k-1)/2$ problems. We again assume nearly equal priors and nearly equal energy on the class means (Assumption 1).

Theorem 9. *Assume that the training set follows a multiclass GMM with noise covariance $\mathbf{\Sigma} = \mathbf{I}_p$ and Assumption 1 holds. Then, there exist constants $c_1, c_2, c_3 > 1$ and $C_1, C_2 > 1$ such that the solutions of the OvA-SVM and MNI are identical with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$ provided that*

$$p > C_1 n \log(kn) + (2n/k) - 1 \qquad \text{and} \qquad p > C_2 n^{1.5} \|\boldsymbol{\mu}\|_2. \tag{115}$$

Proof sketch. Note that the margins of OvO SVM are 1 and -1, hence the proof is similar to the proof of Theorem 7. Recall that in OvO SVM, we solve $k(k-1)/2$ binary problems and each problems has sample size $2n/k$ with high probability. Therefore, compared to OvA SVM which solves k problems each with sample size n, OvO SVM needs less overparameterization to achieve interpolation. Thus the first condition in Equation (109) reduces to $p > C_1 n \log(kn) + (2n/k) - 1$. □

We now derive the classification risk for OvO SVM classifiers. Recall that OvO classification solves $k(k-1)/2$ binary subproblems. Specifically, for each pair of classes, say $(i, j) \in [k] \times [k]$, we train a classifier $\mathbf{w}_{ij} \in \mathbb{R}^p$ and the corresponding decision rule for a fresh sample $\mathbf{x} \in \mathbb{R}^p$ is $\hat{y}_{ij} = \mathrm{sign}(\mathbf{x}^T\hat{\mathbf{w}}_{ij})$. Overall, each class $i \in [k]$ gets a voting score $s_i = \sum_{j \neq i} \mathbf{1}_{\hat{y}_{ij}=+1}$. Thus, the final decision is given by majority rule that *decides the class with the highest score*, i.e. $\arg\max_{i \in [k]} s_i$. Having described the classification process, the total classification error $\mathbb{P}_e$ for balanced classes is given by the conditional error $\mathbb{P}_{e|c}$ given the fresh sample belongs to class c. Without loss of generality, we assume $c = 1$. Formally, $\mathbb{P}_e = \mathbb{P}_{e|1} = \mathbb{P}_{e|1}(s_1 < s_2 \text{ or } s_1 < s_3 \text{ or} \cdots \text{or } s_1 < s_k)$. Under the nearly equal prior and energy assumption, by symmetry and union bound, the conditional classification risk given that true class is 1 can be upper bounded as:

$$\mathbb{P}_{e|1}(s_1 < s_2 \text{ or } s_1 < s_3 \text{ or} \cdots \text{or } s_1 < s_k) \leq \mathbb{P}_{e|1}(s_1 < k-1) = \mathbb{P}_{e|1}(\exists j \ s.t. \ \hat{y}_{1j} \neq 1) \leq (k-1)\mathbb{P}_{e|1}(\hat{y}_{12} \neq 1).$$

Therefore, it suffices to bound $\mathbb{P}_{e|1}(y_{12} \neq 1)$. We can directly apply Theorem 4 with changing k to 2 and n to $2n/k$.

Theorem 10. *Let Assumptions 1 and 2, and the condition in Equation (115) hold. Further assume constants $C_1, C_2, C_3 > 1$ such that $\left(1 - C_1\sqrt{\frac{k}{n}} - \frac{C_2 n}{kp}\right)\|\boldsymbol{\mu}\|_2 > C_3$. Then, there exist additional constants c_1, c_2, c_3 and $C_4 > 1$ such that the OvO SVM solutions satisfies:*

$$\mathbb{P}_{e|c} \leq (k-1)\exp\left(-\|\boldsymbol{\mu}\|_2^2 \frac{\left(\left(1 - C_1\sqrt{\frac{k}{n}} - \frac{C_2 n}{kp}\right)\|\boldsymbol{\mu}\|_2 - C_3\right)^2}{C_4\left(\|\boldsymbol{\mu}\|_2^2 + \frac{kp}{n}\right)}\right) \tag{116}$$

with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$, for every $c \in [k]$. Moreover, the same bound holds for the total classification error $\mathbb{P}_e$.